SIAM J. Math. Anal.
Vol. 31, No. 1, pp. 1–18

BLOW-UP SOLUTIONS FOR A CLASS OF SEMILINEAR ELLIPTIC AND PARABOLIC EQUATIONS[*]

YIHONG DU[†] AND QINGGUANG HUANG[†]

Abstract. We study the asymptotic behavior of the solutions to the problem

$$\begin{cases} u_t - \Delta u &= au - b(x)u^p \quad \text{in} \quad (0,\infty)\times\Omega, \\ \alpha u_\nu + \beta u &= 0 \quad \text{on} \quad (0,\infty)\times\partial\Omega, \\ u(0,.) &= u_0 \quad \text{in} \quad \Omega, \end{cases}$$

where $p > 1$, $b(x) \geq 0$ is continuous and vanishes on the closure of a nontrivial subdomain Ω_0 of $\Omega \subset R^N$. This case can be regarded as a mixture of the well-understood logistic (when $b(x) > 0$ always) and Malthusian (when $b(x) \equiv 0$) models and has attracted much study in recent years. It follows from recent studies that the model behaves like the logistic model if the growth rate a of the species is less than some constant $a_0 > 0$ and it behaves differently from the logistic model once $a \geq a_0$. In this paper, we show that, when $a \geq a_0$, the model behaves like the Malthusian model on part of the domain (i.e., on Ω_0 where b vanishes) and it behaves like the logistic model on the remaining part of the domain. Our study shows that the boundary blow-up problem

$$-\Delta u = au - b(x)u^p \text{ in } \Omega\setminus\overline{\Omega}_0,\ \alpha u_\nu + \beta u = 0 \text{ on } \partial\Omega,\ \ u = \infty \text{ on } \partial\Omega_0$$

plays a key role in understanding the dynamics of our model and that the whole theory can be described by a nice bifurcation picture involving a branch of positive solutions at "infinity."

Key words. logistic equation, boundary blow-up

AMS subject classifications. 35J65, 35K60, 92D25

PII. S0036141099352844

1. Introduction.

In this paper, we consider the semilinear elliptic equation

$$-\Delta u = au - b(x)u^p \text{ in } \Omega,\ \ Bu = 0 \text{ on } \partial\Omega \tag{1.1}$$

and the corresponding parabolic problem

$$\begin{cases} u_t - \Delta u &= au - b(x)u^p \quad \text{in} \quad (0,\infty)\times\Omega, \\ Bu &= 0 \quad \text{on} \quad (0,\infty)\times\partial\Omega, \\ u(0,.) &= u_0 \quad \text{in} \quad \Omega. \end{cases} \tag{1.2}$$

Here a is a real parameter, $b \geq 0$ is in $C^\mu(\bar{\Omega})$, and $p > 1$ is a constant; Ω is a $C^{2+\mu}$ bounded domain in R^N, $N \geq 2$, and the boundary condition is given by

$$Bu = \alpha u_\nu + \beta u,$$

where ν is the unit outward normal to $\partial\Omega$ and either $\alpha = 0, \beta = 1$ (which gives the Dirichlet boundary condition) or $\alpha = 1, \beta \geq 0$ is in $C^{1+\mu}(\partial\Omega)$ (which gives the Neumann or Robin boundary conditions).

Problems (1.1) and (1.2) are basic population models (see, e.g., [Hs]). Problem (1.1) is also related to some prescribed curvature problems in Riemannian geometry

[*] Received by the editors February 22, 1999; accepted for publication (in revised form) April 2, 1999; published electronically November 4, 1999.
http://www.siam.org/journals/sima/31-1/35284.html

[†] School of Mathematical and Computer Sciences, University of New England, Armidale, NSW 2351, Australia (ydu@turing.une.edu.au, qhuang@turing.une.edu.au). The research of Y. Du was partially supported by a grant of the Australian Research Council and a University of New England Research Grant.

(see, e.g., [Ou] and [KW]). We are interested only in positive solutions of (1.1) and (1.2) as these are the solutions which are interesting to us.

If $b(x) > 0 \ \forall \ x \in \overline{\Omega}$, then the equations are known as the logistic equations, and it is well known that (1.1) has a unique positive solution if and only if $a > \lambda_1(\Omega)$, where $\lambda_1(\Omega)$ denotes the first eigenvalue of

$$-\Delta u = \lambda u \text{ in } \Omega, \quad Bu = 0 \text{ on } \ \partial\Omega.$$

Moreover, $u \equiv 0$ attracts all the solutions of (1.2) with admissible nonnegative initial values if $a \leq \lambda_1(\Omega)$, while when $a > \lambda_1(\Omega)$, the unique positive solution of (1.1) attracts all the solutions of (1.2) with admissible nontrivial nonnegative initial values. That is, for any given admissible initial value $u_0 \geq 0$, $u_0 \not\equiv 0$, the unique solution $u(t, x)$ of (1.2) exists for all time $t > 0$, and as $t \to \infty$ it converges to 0 when $a \leq \lambda_1(\Omega)$ and it converges to the unique positive solution $u(x)$ of (1.1) when $a > \lambda_1(\Omega)$.

If $b(x) \equiv 0$, then the equations reduce to the linear Malthusian models with diffusion. It follows from elementary theory of linear parabolic equations (see, e.g., [Fr] or [LSU]) that, as $t \to \infty$, the solution $u(t, x)$ of (1.2) (with nontrivial nonnegative u_0) converges to 0 if $a < \lambda_1(\Omega)$ and it blows up at an exponential rate in t on the whole Ω if $a > \lambda_1(\Omega)$.

We are interested in the degenerate logistic case where $b \geq 0$, $b \not\equiv 0$, but the zero set of b is the closure of some suitably regular nonempty subdomain Ω_0:

$$\overline{\Omega}_0 = \{x \in \overline{\Omega} : b(x) = 0\}.$$

Hence the model becomes a mixture of the logistic and Malthusian models. The assumption that b vanishes on Ω_0 may be interpreted as Ω_0 being an ideal environment, so that the species on Ω_0 has almost no limitation for its population growth. We make this assumption from now on, and let $\lambda_1^D(\Omega_0)$ denote the first eigenvalue of the Dirichlet problem

$$-\Delta u = \lambda u, \quad u|_{\partial\Omega_0} = 0.$$

Under the above assumptions, the elliptic problem (1.1) was studied in [AT], [AG], [Da], [dP], [FKLM], and [Ou]; and [FKLM] also considered the corresponding parabolic problem (1.2). Their results can be summarized as follows (see, e.g., Theorems 3.5 and 3.7 of [FKLM]):

- Equation (1.1) has a positive solution if and only if $a \in (\lambda_1(\Omega), \lambda_1^D(\Omega_0))$. In this case (1.1) has a unique positive solution u_a, $a \to u_a$ is continuous as a map from $(\lambda_1(\Omega), \lambda_1^D(\Omega_0))$ to $C^{2+\mu}(\overline{\Omega})$, and $\|u_a\|_\infty \to \infty$ as $a \to \lambda_1^D(\Omega_0) - 0$.
- For $a \in (\lambda_1(\Omega), \lambda_1^D(\Omega_0))$, the unique positive solution u_a attracts all the solutions of (1.2) with admissible nontrivial nonnegative initial values.
- When $a \leq \lambda_1(\Omega)$, then $u \equiv 0$ attracts all the solutions of (1.2) with admissible nonnegative initial data.
- In the remaining case $a \geq \lambda_1^D(\Omega_0)$, any solution of (1.2) with admissible nontrivial nonnegative initial data blows up in the L^∞-norm as $t \to \infty$: $\lim_{t\to\infty} \|u(t, .)\|_\infty = \infty$.

We remark that $\lambda_1(\Omega) < \lambda_1^D(\Omega_0)$ always holds as $\lambda_1(\Omega) \leq \lambda_1^D(\Omega) < \lambda_1^D(\Omega_0)$.

Note that the above results imply that (1.2) behaves like the logistic model when $a < \lambda_1^D(\Omega_0)$ and it behaves differently from the logistic model when $a \geq \lambda_1^D(\Omega_0)$. It is natural to ask whether it behaves like the Malthusian model in the latter case. This question will be completely answered in this paper. It turns out that our model behaves like the Malthusian model only on part of the domain Ω (in fact, on Ω_0 where b vanishes), and on the rest of the domain, it behaves like the logistic model. Moreover, our results show that the whole theory can be explained by a nice bifurcation picture involving a branch of solutions at "infinity."

Let us now explain our main results in the following. Throughout the paper, we assume that $\overline{\Omega}_0 \subset\subset \Omega$ *is nonempty, connected, and with* $C^{2+\mu}$ *boundary.* The following problem will play an important role:

$$-\Delta u = au - b(x)u^p \text{ in } \Omega \setminus \overline{\Omega}_0,\ Bu = 0 \text{ on } \partial\Omega,\quad u = \infty \text{ on } \partial\Omega_0. \tag{1.3}$$

Here, as usual, $u = \infty$ on $\partial\Omega_0$ means that

$$u(x) \to \infty \text{ as } x \in \Omega \setminus \overline{\Omega}_0 \text{ and } d(x, \partial\Omega_0) \to 0.$$

Now we are ready to state our main results; for simplicity of statement, we have refrained from giving the most general form here.

THEOREM 1.1. (i) *For any* $a \in (-\infty, \infty)$, *problem* (1.3) *has a minimal positive solution* $\underline{U}_a$ *and a maximal positive solution* $\overline{U}_a$.

(ii) *If there exist positive constants* α *and* c *such that* $\lim_{d(x,\Omega_0)\to 0} b(x)/[d(x, \Omega_0)]^\alpha = c$, *then* (1.3) *has a unique positive solution.*

THEOREM 1.2. *Let* $a_0 = \lambda_1^D(\Omega_0)$. *Then*

(i) *for any fixed* $x \in \overline{\Omega}_0$, $u_a(x) \to \infty$ *as* $a \nearrow a_0$;

(ii) *for any fixed* $x \in \overline{\Omega} \setminus \overline{\Omega}_0$, $u_a(x) \to \underline{U}_{a_0}(x)$ *as* $a \nearrow a_0$.

THEOREM 1.3. *Let* $a \geq a_0 = \lambda_1^D(\Omega_0)$. *Then for any given admissible nontrivial nonnegative initial value* u_0, *the unique solution* $u(t, x)$ *of* (1.2) *satisfies*

(i) *for any fixed* $x \in \overline{\Omega}_0$, $u(t, x) \to \infty$ *as* $t \to \infty$;

(ii) *for any fixed* $x \in \overline{\Omega} \setminus \overline{\Omega}_0$, $\overline{\lim}_{t\to\infty} u(t, x) \leq \overline{U}_a(x)$; $\underline{\lim}_{t\to\infty} u(t, x) \geq \underline{U}_a(x)$;

(iii) *if* (1.3) *has a unique positive solution denoted as* U_a, *then for any fixed* $x \in \overline{\Omega} \setminus \overline{\Omega}_0$, $\lim_{t\to\infty} u(t, x) = U_a(x)$.

When the solution of (1.3) is unique (as in case (ii), Theorem 1.1), then the above results give the following nice bifurcation picture for (1.1) and (1.2):

Denote $\tilde{U}_a(x) = U_a(x)$ *when* $a \in \Omega \setminus \overline{\Omega}_0$ *and* $\tilde{U}_a(x) = \infty$ *when* $x \in \overline{\Omega}_0$; *and regard* $\tilde{U}_a$ *as a solution of* (1.1) *at "infinity." Then the unique positive solution branch*

$$\Sigma = \{(a, u_a) : \lambda_1(\Omega) < a < \lambda_1^D(\Omega_0)\}$$

bifurcates from the branch of trivial solutions

$$\Sigma_0 = \{(a, 0) : -\infty < a < \infty\}$$

at $a = \lambda_1(\Omega)$ *and it joins the branch of positive solutions at "infinity,"*

$$\Sigma_\infty = \{(a, \tilde{U}_a) : -\infty < a < \infty\},$$

at $a = \lambda_1^D(\Omega_0)$. *Moreover, when* $a \leq \lambda_1(\Omega)$, *the trivial solution on* Σ_0 *is globally attractive for* (1.2), *when* $\lambda_1(\Omega) < a < \lambda_1^D(\Omega_0)$, *the positive solution on* Σ *is globally attractive, and when* $a \geq \lambda_1^D(\Omega_0)$, *the solution at "infinity" on* Σ_∞ *is globally attractive.*

Remark 1.4. (i) Our results remain valid if $b(x)$ vanishes on the closure of a finite number of disjoint subdomains $\Omega_1, \ldots, \Omega_k$, all with $C^{2+\mu}$ boundary; the statements of our results being modified accordingly.

(ii) With a little more effort, our arguments (except the proofs of Theorem 2.8 and Proposition 4.4) can be carried out when Δ is replaced by a general self-adjoint second-order strongly elliptic operator and the term u^p replaced by a more general nonlinear function of the similar type.

(iii) Many difficulties in our arguments come from the fact that $b(x)$ vanishes on Ω_0 in a continuous fashion. If we allow $b(x)$ to be discontinuous on $\partial\Omega_0$ so that $b(x) \geq c_0 > 0$ on $\Omega \setminus \Omega_0$, then it is much easier to establish similar or better results (and with less regularity on $\partial\Omega_0$). For example, in this case, (1.3) has a unique positive solution (by a variant of [MV]). We suspect that the uniqueness result for (1.3) always holds.

(iv) Our techniques in this paper rely heavily on the assumption that $\Omega_0 \subset\subset \Omega$ and do not work for the case $\partial\Omega_0 \cap \partial\Omega \neq \emptyset$. This latter case is rather difficult and is discussed in [DG].

(v) As a by-product of our results, we find that the sufficient condition on the asymptotic behavior of the first eigenvalues given in [FKLM] is far from necessary. To be more specific, Theorem 2.4 of [FKLM] shows that

$$\lambda_1(-\Delta + q_k, \Omega) \to \lambda_1^D(\Omega_0)$$

if q_k is a sequence of increasing nonnegative functions in $C^\mu(\overline{\Omega})$ satisfying

$$q_k \equiv 0 \text{ on } \Omega_0, \quad q_k(x) \to \infty \text{ uniformly on any compact set } K \subset \overline{\Omega} \setminus \overline{\Omega}_0.$$

However, if we take $q_k(x) = b(x)u_{a_k}^{p-1}(x)$, where $a_k < a_0 = \lambda_1^D(\Omega_0)$ and $a_k \to a_0$, then by Theorem 3.6,

$$q_k(x) \to b(x)\underline{U}_{a_0}^{p-1}(x) < \infty \text{ uniformly on any compact set } K \subset \overline{\Omega} \setminus \overline{\Omega}_0,\ q_k \equiv 0 \text{ on } \Omega_0,$$

but we still have, from the equations for u_{a_k}, that

$$\lambda_1(-\Delta + q_k, \Omega) = a_k \to a_0 = \lambda_1^D(\Omega_0).$$

Finally, we would like to mention that if $b(x)$ changes sign on Ω, then it is known (see, e.g., [AT2] and [BCN]) that there exists $a^* > \lambda_1(\Omega)$ such that (1.1) has a positive solution if and only if

(i) $a < \lambda_1(\Omega)$ or

(ii) $a \in [\lambda_1(\Omega), a^*]$ and $\int_\Omega b(x)\phi(x) < 0$, where $\phi > 0$ is an eigenfunction corresponding to $\lambda_1(\Omega)$.

Some multiplicity results can also be found in [AT2].

The rest of the paper is organized as follows. Section 2 is devoted to the boundary blow-up problem (1.3), where we use various upper and lower solution arguments to prove the existence and uniqueness of positive solutions to (1.3). In section 3, we study the asymptotic behavior of the positive solution u_a of (1.1) as $a \nearrow a_0$. A crucial step here is to show that the solutions blow up on $\partial\Omega_0$. In section 4, we make use of the results obtained in the previous sections and discuss how the solution of (1.2) blows up as $t \to \infty$.

After this paper was submitted for publication, we learned of the works [GGLS] and [LS], where some related problems were discussed and interesting numerical simulations were presented. Our method in this paper can be used to improve some of the results on the blow-up behavior in [GGLS] and [LS]. More precisely, in [LS, Theorem 4.3] (see also [GGLS, Corollary 3.3]), through analysis on the first variation of the principal eigenvalues, it is proved that the stationary solution blows up on the boundary Γ of $\{b(x)>0\}$ as a approaches $\lambda_1^D(\{b(x)=0\})$, provided that

$$b(x)=o(d(x,\Gamma)) \text{ as } d(x,\Gamma)\to 0.$$

The method in section 3 of the present paper shows that the above condition is unnecessary.

2. The boundary blow-up problem (1.3). In this section, we study in detail problem (1.3). Boundary blow-up problems similar to (1.3) arise naturally from a number of different areas and have a long history. Considerable amounts of study have been attracted by such problems. We mention only [BM], [LM], [LN], and [MV]; many other works can be found from the references of these papers. One main difference, which posses technical difficulties, of our problem (1.3) from the previous ones is that our function $b(x)$ vanishes on $\partial\Omega_0$.

We start with an interesting comparison result which will be used frequently later (actually this result is a little more than enough for our later use).

LEMMA 2.1. *If $u_1,u_2\in C^2(\overline{\Omega}\setminus\overline{\Omega}_0)$ are both positive in $\Omega\setminus\overline{\Omega}_0$ and*

$$\Delta u_1+au_1-b(x)u_1^p\le 0\le \Delta u_2+au_2-b(x)u_2^p \textit{ in } \Omega\setminus\overline{\Omega}_0, \tag{2.1}$$

$$Bu_1\ge Bu_2 \textit{ on } \partial\Omega;\quad \overline{\lim}_{d(x,\partial\Omega_0)\to 0}(u_2-u_1)\le 0,$$

then $u_1\ge u_2$ on $\overline{\Omega}\setminus\overline{\Omega}_0$.

Proof. We shall use a variant of a method used in the proof of Lemma 1.1 in [MV] which goes back to [BBL].

We consider only the case that the boundary operator B in our problem is of Neumann or Robin type, as the Dirichlet case can be proved in exactly the same manner as in [MV]. Let w_1,w_2 be nonnegative C^2 functions on $\overline{\Omega}\setminus\Omega_0$ vanishing near $\partial\Omega_0$. Using (2.1) and applying integration by parts and subtraction we easily obtain

$$\begin{aligned}&-\int_{\tilde\Omega}[\nabla u_2\nabla w_2-\nabla u_1\nabla w_1]-\beta\int_{\partial\Omega}(u_2w_2-u_1w_1)\\ &\qquad\ge\int_{\tilde\Omega}b(x)(u_2^pw_2-u_1^pw_1)+a\int_{\tilde\Omega}(u_1w_1-u_2w_2),\end{aligned} \tag{2.2}$$

where $\tilde\Omega=\Omega\setminus\overline{\Omega}_0$.

Let $\epsilon_1>\epsilon_2>0$ and denote

$$v_i=(u_i+\epsilon_i)^{-1}[(u_2+\epsilon_2)^2-(u_1+\epsilon_1)^2]^+,\qquad i=1,2.$$

Since v_i can be approximated arbitrarily closely in the $W^{1,2}\cap L^\infty$ norm on $\overline{\Omega}\setminus\Omega_0$ by C^2 functions vanishing near $\partial\Omega_0$ (this is easily seen to be possible if, say, one extends v_i continuously across $\partial\Omega_0$ first and then uses mollifiers), we see that (2.2) holds when w_i is replaced by v_i. Denote

$$\Omega_+(\epsilon_1,\epsilon_2)=\{x\in\tilde\Omega: u_2(x)+\epsilon_2>u_1(x)+\epsilon_1\},$$

and note that the integrands of $\int_{\tilde\Omega}$ in (2.2) (with $w_i=v_i$) vanish outside this set. The first integral on the left-hand side of (2.2) equals

$$-\int_{\Omega_+(\epsilon_1,\epsilon_2)}\left(\left|\nabla u_2-\frac{u_2+\epsilon_2}{u_1+\epsilon_1}\nabla u_1\right|^2+\left|\nabla u_1-\frac{u_1+\epsilon_1}{u_2+\epsilon_2}\nabla u_2\right|^2\right),$$

which is nonpositive. On the other hand, as $0 < \epsilon_2 < \epsilon_1 \to 0$, the first term on the right-hand side of (2.2) converges to

$$\int_{\Omega_+(0,0)} b(x)(u_2^{p-1} - u_1^{p-1})(u_2^2 - u_1^2)$$

and the other two remaining terms in (2.2) converge to 0. Therefore we would have a contradiction unless $\Omega_+(0,0)$ has measure 0, i.e., $u_1 \geq u_2$ on $\tilde{\Omega}$. □

Remark 2.2. The conclusion of Lemma 2.1 holds under much weaker conditions; see, e.g., Lemma 1.1 in [MV], where Ω_0 is empty.

Lemma 2.3. *For any given positive function $\phi \in C^{2+\mu}(\partial\Omega_0)$ and $a \in (-\infty, \infty)$, the problem*

$$-\Delta u = au - b(x)u^p \text{ in } \Omega \setminus \overline{\Omega}_0;\ u|_{\partial\Omega_0} = \phi,\ Bu|_{\partial\Omega} = 0 \tag{2.3}$$

has a unique positive solution.

Proof. Let $a^* = \max\{\lambda_1(\Omega), a\}$. Then choose a smooth nonnegative function $b^*(x)$ such that $b^*(x) \leq b(x)$ on $\Omega \setminus \Omega_0$ and $\Omega_0^* \equiv \{x \in \overline{\Omega} : b^*(x) = 0\}$ has small volume so that $\lambda_1^D(\Omega_0^*) > a^*$. Then, by Theorem 2 of [Ou] (see also [AT], [dP], or [FKLM] for the Robin boundary conditions), there is a unique positive solution u^* for the problem

$$-\Delta u = a^*u - b^*(x)u^p, \quad Bu|_{\partial\Omega} = 0.$$

Choose a large constant $M > 1$ such that $Mu^* > \phi$ on $\partial\Omega_0$. Then it is easily checked that Mu^* is a supersolution to (2.3). Clearly $u \equiv 0$ is a subsolution to (2.3). Therefore, (2.3) has at least one positive solution (see, for example, [St]). By Lemma 2.1, there is at most one positive solution. Hence (2.3) has a unique positive solution. □

Theorem 2.4. *For any $a \in (-\infty, \infty)$,* (1.3) *has a minimal positive solution $\underline{U}_a$ and a maximal positive solution $\overline{U}_a$ in the sense that any positive solution u of* (1.3) *satisfies $\underline{U}_a(x) \leq u(x) \leq \overline{U}_a(x)$.*

Proof. Let $u_c(x)$ denote the unique positive solution of (2.3) with $\phi(x) \equiv c > 0$. We know from Lemma 2.1 that $c \to u_c(x)$ is increasing. If we can show that $u_c(x)$ is bounded from above by some function $V(x)$ which is uniformly bounded on all compact subsets of $\overline{\Omega} \setminus \overline{\Omega}_0$, then a simple regularity and compactness argument shows that $u_\infty(x) \equiv \lim_{c\to\infty} u_c(x)$ is a solution of (1.3). We now set out to find such a function V.

We can find a nonnegative function $b^*(x)$ in $C^2(\overline{\Omega} \setminus \Omega_0)$ such that

$$0 < b^*(x) \leq b(x) \quad \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0.$$

Such a function is easy to construct for x bounded away from $\partial\Omega_0$. For x satisfying $0 < d(x, \Omega_0) < \delta$, where $\delta > 0$ is small such that $x \to d(x, \Omega_0)$ is C^2, we can define

$$b^*(x) = f(d(x, \Omega_0)), \quad \text{where } f(\eta) = \int_0^\eta \int_0^\xi [\min_{d(x,\Omega_0)\geq s} b(x)] ds d\xi.$$

We now fix a $c_0 > 0$ and define $V^*(x)$ such that

(i) $V^*(x) = u_{c_0}(x)$ for $x \in \overline{\Omega}$ and near $\partial\Omega$;

(ii) $V^*(x) = [b^*(x)]^\beta,\ \beta = 3/(1-p)$ for x satisfying $0 < d(x, \Omega_0) < \delta$;

(iii) V^* is C^2 and positive on $\overline{\Omega} \setminus \overline{\Omega}_0$.

We show that, for all large constant $M > 0$, $V(x) = MV^*(x)$ meets our requirement specified earlier.

Since

$$BV(x) = BMu_{c_0}(x) = 0 \ \forall x \in \partial\Omega \text{ and } \lim_{d(x,\Omega_0)\to 0}[u_c(x) - V(x)] = -\infty < 0,$$

by Lemma 2.1, we will have $u_c(x) \le V(x) \ \forall \ \ x \in \overline{\Omega} \setminus \overline{\Omega}_0$ if we can show that

$$-\Delta V \ge aV - b(x)V^p \quad \forall x \in \Omega \setminus \overline{\Omega}_0. \tag{2.4}$$

For x satisfying $0 < d(x, \Omega_0) < \delta$, a direct calculation gives

$$-\Delta V - aV + b(x)V^p$$

$$= -\beta M[b^*(x)]^{\beta-1}\Delta b^*(x) - \beta(\beta-1)M[b^*(x)]^{\beta-2}|\nabla b^*(x)|^2 - aM[b^*(x)]^{\beta} + M^p b(x)[b^*(x)]^{p\beta}$$

$$\ge M[b^*(x)]^{p\beta+1}\{-\beta b^*(x)\Delta b^*(x) - \beta(\beta-1)|\nabla b^*(x)|^2 - a[b^*(x)]^2 + M^{p-1}\} > 0$$

$\forall$ large $M > 0$. For $x \in \overline{\Omega}$ satisfying $d(x, \Omega_0) \ge \delta$,

$$-\Delta V - aV + b(x)V^p = M[-\Delta V^* - aV^* + b(x)M^{p-1}V^{*p}] \ge 0$$

$\forall$ large M. Hence (2.4) is always satisfied if M is large.

By Lemma 2.1, any solution u of (1.3) satisfies $u \ge u_c$ for every $c > 0$. Hence $\underline{U} \equiv \lim_{c\to\infty} u_c \le u$ and $\underline{U}$ is a minimal positive solution of (1.3).

To show the existence of a maximal solution for (1.3), we consider the problem

$$-\Delta u = au - b(x)u^p \text{ in } \Omega \setminus \overline{\Omega}_n, \quad u|_{\partial\Omega_n} = \infty, \ Bu|_{\partial\Omega} = 0,$$

where $\Omega_n \equiv \{x \in \Omega : d(x, \Omega_0) < 1/n\}$.

A similar but simpler (since $b(x) > 0$ on $\overline{\Omega}\setminus\Omega_n$) argument shows that this problem has a minimal positive solution u_n. Using Lemma 2.1, we see that for any positive solution u of (1.3), $u_n \ge u_{n+1} \ge u$ on $\Omega \setminus \overline{\Omega}_n$. Hence $\overline{U}(x) = \lim_{n\to\infty} u_n(x) \ge u(x)$. But one easily sees that $\overline{U}$ is a positive solution of (1.3). Hence it is a maximal positive solution. This finishes the proof of Theorem 2.4. □

Remark 2.5. To show the convergence of u_c as $c \to \infty$, we actually need to find only, for each open set O with $\overline{O} \subset \overline{\Omega} \setminus \overline{\Omega}_0$, a function V_O such that $V_O \ge u_c$ on O for all c. This is much easier to do and requires no regularity on $\partial\Omega_0$. In fact, in a number of places in this paper, our assumption on the regularity of the domain is more than necessary. Here we have kept the above longer proof because it shows that if $b(x)$ is C^2 near $\partial\Omega_0$, then $\underline{U}_a(x) \le M[b(x)]^{3/(1-p)}$ for x near $\partial\Omega_0$.

Next we discuss the uniqueness of the positive solutions of (1.3). The following result will be useful in our proof of the uniqueness result.

LEMMA 2.6. *Let $b^* \in C^{\mu}(\overline{\Omega})$ be such that $b^*(x) > 0 \ \forall x \in \overline{\Omega} \setminus \partial\Omega_0$. Then, for any $a \in (-\infty, \infty)$, the problem*

$$-\Delta u = au - b^*(x)u^p \ \textit{in } \Omega, \quad u|_{\partial\Omega} = \infty \tag{2.5}$$

has at least one positive solution.

Proof. Choose subdomains Ω_1 and Ω_2 such that

$$\overline{\Omega}_0 \subset\subset \Omega_1, \ \overline{\Omega}_1 \subset\subset \Omega_2, \ \overline{\Omega}_2 \subset\subset \Omega.$$

By [MV], the problem

$$-\Delta u = au - b^*(x)u^p \text{ in } \Omega \setminus \overline{\Omega}_1, \quad u|_{\partial\Omega} = \infty, \; u|_{\partial\Omega_1} = \infty$$

has a positive solution $u_1(x)$.

Let $a^* > \max\{a, \lambda_1(\Omega)\}$. Then the problem

$$-\Delta u = a^* u - b^*(x)u^p \text{ in } \Omega, \quad u|_{\partial\Omega} = 0$$

has a positive solution $u_2(x)$ (see, e.g., [dP]).

Now we define $u^* \in C^2(\overline{\Omega})$ such that

(i) $u^*(x) = u_1(x)$ for $x \in \Omega \setminus \Omega_2$;

(ii) $u^*(x) = u_2(x)$ for $x \in \Omega_1$;

(iii) $u^*(x) > 0$ on Ω.

Then it is easily checked that for all large constant $M > 0$, $U = Mu^*$ satisfies

$$-\Delta U(x) \geq aU(x) - b^*(x)U^p(x) \;\; \forall x \in \Omega.$$

A super- and subsolution argument, together with the use of Lemma 2.1, shows that the problem

$$-\Delta u = au - b^*(x)u^p \text{ in } \Omega, \quad u|_{\partial\Omega} = n$$

has a unique positive solution u_n and $u_n \leq u_{n+1} \leq U$ on Ω. Hence $u_\infty(x) \equiv \lim_{n\to\infty} u_n(x)$ exists and is a positive solution of (2.5). This finishes the proof of Lemma 2.6. □

Remark 2.7. It is easy to show that (2.5) has a unique positive solution (see [MV]). Lemma 2.6 generalizes earlier results of this type (see, e.g., [BM], [MV]), where $b^* > 0$ on the entire domain is required. One easily sees from the above proof how this result can be generalized to the case that b^* vanishes on the closure of a subdomain of Ω.

THEOREM 2.8. *Denote $d(x) = d(x, \Omega_0)$. Suppose there exist positive constants α and c such that*

$$\lim_{d(x)\to 0} b(x)/[d(x)]^\alpha = c.$$

Then for any $a \in (-\infty, \infty)$, problem (1.3) *has a unique positive solution U_a. Moreover,*

$$\lim_{d(x)\to 0} [d(x)]^{(\alpha+2)/(p-1)} U_a(x) = \Big(\frac{(2+\alpha)(1+\alpha+p)}{c(p-1)^2}\Big)^{1/(p-1)}.$$

Proof. Given any small $\epsilon > 0$, we fix a $\delta > 0$ small such that

(i) $d(x)$ is C^2 $\forall$ x satisfying $0 < d(x) < 2\delta$;

(ii) $|\frac{2+\alpha}{p-1}\Delta d(x)s - as^2| < \epsilon$ $\forall$ $s \in [0, 2\delta]$ and x satisfying $0 < d(x) < 2\delta$;

(iii) $(c-\epsilon)d(x)^\alpha \leq b(x) \leq (c+\epsilon)d(x)^\alpha$ $\forall$ x with $0 < d(x) < 2\delta$.

Denote $\beta = -\frac{2+\alpha}{p-1}$, and let

$$\underline{\xi} = \Big(\frac{\beta(\beta-1)-\epsilon}{c+\epsilon}\Big)^{1/(p-1)}, \quad \overline{\xi} = \Big(\frac{\beta(\beta-1)+\epsilon}{c-\epsilon}\Big)^{1/(p-1)}, \tag{2.6}$$

and for $\sigma \in (0, \delta)$, define

$$\underline{v}_\sigma = [d(x)+\sigma]^\beta \underline{\xi}, \quad \overline{v}_\sigma = [d(x)-\sigma]^\beta \overline{\xi}.$$

Since $|\nabla d(x)| \equiv 1$ and $b(x) \geq (c-\epsilon)[d(x)-\sigma]^\alpha$ when $\sigma < d(x) < 2\delta$, we easily obtain

$$-\Delta \overline{v}_\sigma - a\overline{v}_\sigma + b(x)(\overline{v}_\sigma)^p$$

$$= \overline{\xi}\Big\{-\beta[d(x)-\sigma]^{\beta-1}\Delta d(x) - \beta(\beta-1)[d(x)-\sigma]^{\beta-2} - a[d(x)-\sigma]^\beta + b(x)[d(x)-\sigma]^{\beta p}\overline{\xi}^{p-1}\Big\}$$

$$\geq \overline{\xi}[d(x)-\sigma]^{\beta-2}\Big\{-\beta\Delta d(x)[d(x)-\sigma] - a[d(x)-\sigma]^2 + \epsilon\Big\}$$

$$\geq 0 \quad \forall\ x \text{ satisfying } \sigma < d(x) < 2\delta.$$

Similarly, since $b(x) < (c+\epsilon)[d(x)+\sigma]^\alpha$ when $d(x)+\sigma < 2\delta$,

$$-\Delta \underline{v}_\sigma - a\underline{v}_\sigma + b(x)(\underline{v}_\sigma)^p$$

$$\leq \underline{\xi}[d(x)+\sigma]^{\beta-2}\Big\{-\beta\Delta d(x)[d(x)+\sigma] - a[d(x)+\sigma]^2 - \epsilon\Big\}$$

$$\leq 0 \quad \forall\ x \text{ satisfying } d(x)+\sigma < 2\delta.$$

Let w be the positive solution of (2.5) with Ω replaced by $\Omega_\delta \equiv \{x \in \Omega : d(x,\Omega_0) < \delta\}$ and $b^* \in C^\mu(\overline{\Omega_\delta})$ satisfying $b^*(x) = b(x)$ for $x \in \overline{\Omega_\delta} \setminus \Omega_0$ and $b^*(x) > 0$ for $x \in \Omega_0$.

Suppose that u is any positive solution of (1.3). Then one easily checks that $v = u + w$ satisfies

$$-\Delta v \geq av - b(x)v^p \quad \text{in } \Omega_\delta \setminus \overline{\Omega_0}.$$

Since

$$v|_{\partial\Omega_0} = \infty > \underline{v}_\sigma|_{\partial\Omega_0} \text{ and } v|_{\partial\Omega_\delta} = \infty > \underline{v}_\sigma|_{\partial\Omega_\delta},$$

by Lemma 2.1,

$$u + w \geq \underline{v}_\sigma \ \text{ on } \Omega_\delta \setminus \overline{\Omega}_0. \tag{2.7}$$

Similarly,

$$\overline{v}_\sigma + w \geq u \ \text{ on } \ \Omega_\delta \setminus \overline{\Omega}_\sigma. \tag{2.8}$$

Letting $\sigma \to 0$ in (2.7) and (2.8), we deduce

$$d(x)^\beta\overline{\xi} + 2w \geq u + w \geq d(x)^\beta\underline{\xi} \ \ \forall x \in \Omega_\delta \setminus \overline{\Omega}_0.$$

Since w is uniformly bounded on $\partial\Omega_0$, it follows that

$$\underline{\xi} \leq \underline{\lim}_{d(x)\to 0} d(x)^{-\beta}u(x) \leq \overline{\lim}_{d(x)\to 0} d(x)^{-\beta}u(x) \leq \overline{\xi}. \tag{2.9}$$

Recalling (2.6) and letting $\epsilon \to 0$ in (2.9), we obtain

$$\lim_{d(x)\to 0} d(x)^{-\beta}u(x) = \Big(\frac{\beta(\beta-1)}{c}\Big)^{1/(p-1)} = \Big(\frac{(2+\alpha)(1+\alpha+p)}{c(p-1)^2}\Big)^{1/(p-1)}. \tag{2.10}$$

Suppose now u_1 and u_2 are two positive solutions of (1.3). By (2.10), for any $\epsilon > 0$,

$$\lim_{d(x)\to 0}[u_1(x) - (1+\epsilon)u_2(x)] = -\infty, \lim_{d(x)\to 0}[u_2(x) - (1+\epsilon)u_1(x)] = -\infty.$$

Let us denote $w_i = (1+\epsilon)u_i$, $i = 1, 2$. Clearly,

$$-\Delta w_i \geq a w_i - b(x) w_i^p \text{ in } \Omega \setminus \overline{\Omega}_0, \quad B w_i = 0 \text{ on } \partial\Omega.$$

Hence we can use Lemma 2.1 to conclude that

$$u_1(x) \leq (1+\epsilon)u_2(x), \quad u_2(x) \leq (1+\epsilon)u_1(x) \quad \forall x \in \Omega \setminus \overline{\Omega}_0.$$

Letting $\epsilon \to 0$, we obtain $u_1 \equiv u_2$. This finishes the proof of Theorem 2.8. □

Remark 2.9. From the above proof, we easily see that if $b(x) > 0$ on $\overline{\Omega} \setminus \Omega_0$ and if the condition of Theorem 2.8 holds with $\alpha = 0$, then the uniqueness conclusion and the asymptotic formula near $\partial\Omega_0$ (with $\alpha = 0$) are also valid. In fact, it is easy to show by a simple variant of the techniques in [MV] that when $b(x) > 0$ on $\overline{\Omega} \setminus \Omega_0$, (1.3) always has a unique positive solution. We suspect that, even in our case where $b(x)$ vanishes on $\partial\Omega_0$, the condition $\lim_{d(x)\to 0} b(x)/[d(x)]^\alpha = c$ in Theorem 2.8 is unnecessary for uniqueness.

By using Lemma 2.1 and a simple compactness argument, we deduce easily the following result.

Corollary 2.10. *Suppose that* (1.3) *has a unique positive solution* U_a. *Then*

(i) $a \to U_a$ *is continuous as a map from* $(-\infty, \infty)$ *to* $C^{2+\mu}(K)$ *for any compact set* $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$;

(ii) *for any fixed* $x \in \Omega \setminus \overline{\Omega}_0$, $a \to U_a(x)$ *is strictly increasing.*

3. Blow-up solutions of problem (1.1). In this section, we study the blow-up solutions of (1.1) as $a \to a_0$. It is well known that $a \to u_a(x)$ is strictly increasing for $a \in (\lambda_1(\Omega), a_0)$. Therefore, to study the behavior of $\lim_{a\to a_0} u_a(x)$, it suffices to study this limit when $a \to a_0$ is replaced by some sequence $a_n \to a_0$. To this end, we let

$$\Omega_n = \{x \in \Omega : d(x, \Omega_0) < 1/n\}.$$

Without loss of generality, we may assume that $\Omega_n \subset\subset \Omega$ for all $n \geq 1$. Let $a_n = \lambda_1^D(\Omega_n)$. Then we have

$$\lambda_1(\Omega) < a_n < a_0, \quad a_n \to a_0 \text{ as } n \to \infty.$$

We denote $u_n = u_{a_n}$, i.e.,

$$-\Delta u_n = a_n u_n - b(x) u_n^p \text{ in } \Omega, \quad B u_n = 0 \text{ on } \partial\Omega.$$

Lemma 3.1. $\lim_{n\to\infty} u_n(x) = \infty$ *uniformly for* x *in any compact set* $K \subset\subset \Omega_0$.

Proof. Let $\phi_0 \geq 0$ with $\|\phi_0\|_\infty = 1$ be the eigenfunction corresponding to $a_0 = \lambda_1^D(\Omega_0)$,

$$-\Delta\phi_0 = a_0\phi_0, \quad \phi_0|_{\partial\Omega_0} = 0,$$

and let

$$\alpha_0 = \inf_{x\in\Omega_0} u_1(x), \quad \beta_0 = \min_{x\in K} \phi_0(x).$$

Clearly,

$$\alpha_0 > 0,\ \beta_0 > 0,\quad u_n(x) \ge u_1(x) \ge \alpha_0\ \forall n \ge 1, x \in \Omega_0. \tag{3.1}$$

Given any large number $M > 0$, we can find a domain K^* satisfying $K \subset K^* \subset\subset \Omega_0$ such that

$$\phi_0(x) < \alpha_0\beta_0/(2M)\ \forall x \in \partial K^*. \tag{3.2}$$

By a standard interior regularity argument (see, e.g., the proof of Theorem 2.1 in [BNV]), $\phi_n \to \phi_0$ uniformly on K^*, where ϕ_n is given by

$$-\Delta\phi_n = a_n\phi_n,\quad \phi_n|_{\partial\Omega_n} = 0,\ \phi_n \ge 0,\ \|\phi_n\|_\infty = 1.$$

Thus, by (3.2) and the definition of β_0, ∀ large n,

$$(M/\beta_0)\phi_n(x) < \alpha_0\ \forall x \in \partial K^*;\quad (M/\beta_0)\phi_n(x) > M/2\ \forall x \in K. \tag{3.3}$$

Recall that $b(x) = 0$ on K^*. Hence u_n and $(M/\beta_0)\phi_n$ satisfy the same equation $-\Delta u = a_n u$. It now follows from (3.1) and (3.3) that $(M/\beta_0)\phi_n$ and u_n are, respectively, sub- and supersolutions of the problem

$$-\Delta u = a_n u \text{ in } K^*,\quad u|_{\partial K^*} = \alpha_0.$$

As $a_n < a_0 < \lambda_1^D(K^*)$, it follows from the maximum principle that ∀ large n,

$$u_n(x) \ge (M/\beta_0)\phi_n(x) \ge M/2\ \ \forall x \in K \subset K^*.$$

Since $M > 0$ is arbitrary, this shows $\lim_{n\to\infty} u_n(x) = \infty$ uniformly in K. □

Remark 3.2. Lemma 3.1 is related to Theorem 3 of [dP].

Since $\partial\Omega_0$ is $C^{2+\mu}$, it satisfies a uniform interior ball condition: There exists $R > 0$ such that for any $x \in \partial\Omega_0$, there is a ball B_x of radius R such that $B_x \subset \overline{\Omega}_0$ and $B_x \cap \partial\Omega_0 = \{x\}$.

Lemma 3.3. *Let $x_n \in \partial\Omega_0$ be such that*

$$u_n(x_n) = \min_{x\in\partial\Omega_0} u_n(x).$$

If $\{u_n(x_n)\}$ is bounded, then we can find a constant $\sigma > 0$ and a sequence $c_n \to \infty$ such that

$$u_n(x) \ge u_n(x_n) + c_n\psi(x) \quad \textit{whenever } R/2 \le |x - y_n| \le R, \tag{3.4}$$

where $\psi(x) = e^{-\sigma|x-y_n|^2} - e^{-\sigma R^2}$ and y_n is the center of the ball B_{x_n}.

Proof. A simple calculation gives

$$\Delta\psi + a_n\psi = (4\sigma^2|x - y_n|^2 - 2N\sigma + a_n)e^{-\sigma|x-y_n|^2} - a_n e^{-\sigma R^2}.$$

We can choose a large $\sigma > 0$ such that

$$-\Delta\psi(x) \le a_n\psi(x)\ \ \forall x \in B_{x_n} \setminus B_{R/2}(y_n),$$

where $B_{R/2}(y_n) = \{x \in R^N : |x - y_n| < R/2\}$.

Choose a compact set $K \subset\subset \Omega_0$ such that $K \supset \cup_{n=1}^{\infty} B_{R/2}(y_n)$. By Lemma 3.1 and the assumption that $\{u_n(x_n)\}$ is bounded, we can find a sequence $c_n \to \infty$ such that

$$u_n(x) \geq u_n(x_n) + c_n(e^{-\sigma R^2/4} - e^{-\sigma R^2}) \;\; \forall x \in B_{R/2}(y_n) \subset K.$$

On the other hand, since $a_n < a_0$, by the maximum principle, $u_n(x) \geq u_n(x_n)$ $\forall x \in \Omega_0$. In particular, $u_n(x) \geq u_n(x_n)$ on ∂B_{x_n}. Thus we see that u_n is a supersolution to the problem

$$\text{(3.5)} \qquad \begin{cases} -\Delta u = a_n u \text{ in } B_{x_n} \setminus \overline{B}_{R/2}(y_n), \\ u|_{\partial B_{x_n}} = u_n(x_n), \; u|_{\partial B_{R/2}(y_n)} = u_n(x_n) + c_n(e^{-\sigma R^2/4} - e^{-\sigma R^2}). \end{cases}$$

But clearly, $u_n(x_n) + c_n\psi(x)$ is a subsolution to (3.5). Hence, since $a_n < a_0 < \lambda_1^D(B_{x_n} \setminus \overline{B}_{R/2}(y_n))$, by the maximum principle,

$$u_n(x) \geq u_n(x_n) + c_n\psi(x) \;\; \text{whenever } R/2 \leq |x - y_n| \leq R,$$

as required. □

LEMMA 3.4. $\lim_{n\to\infty} u_n(x) = \infty$ *uniformly on* $\overline{\Omega}_0$.

Proof. By the maximum principle, it suffices to show that

$$u_n(x_n) = \min_{x\in\partial\Omega_0} u_n(x) \to \infty.$$

We argue indirectly. Suppose that this is not true. Then by passing to a subsequence, we may assume that $\{u_n(x_n)\}$ is bounded: $u_n(x_n) \leq C$ $\forall$ n.

Clearly, u_n is a supersolution to

$$\text{(3.6)} \qquad -\Delta u = a_n u - b(x)u^p \text{ in } \Omega \setminus \overline{\Omega}_0; u|_{\partial\Omega_0} = u_n(x_n), \; Bu|_{\partial\Omega} = 0.$$

By Lemma 2.3, (3.6) has a unique positive solution, which we denote as v_n. By Lemma 2.1, we deduce $u_n \geq v_n$ on $\Omega \setminus \Omega_0$. Replacing $u_n(x_n)$ in (3.6) by its upper bound C, we similarly obtain a unique positive solution V of (3.6) and by Lemma 2.1, $v_n \leq V$ on $\Omega\setminus\Omega_0$. In particular, $\|v_n\|_{L^\infty(\Omega\setminus\Omega_0)}$ is bounded. Then the L^p-estimates and the Sobolev imbedding theorems (see [GT]) imply that $\{v_n\}$ is bounded in $C^1(\overline{\Omega}\setminus\Omega_0)$. In particular, $|\nabla v_n(x_n)|$ is bounded. Since

$$u_n(x) \geq v_n(x) \;\; \forall x \in \Omega \setminus \Omega_0 \quad \text{and } u_n(x_n) = v_n(x_n),$$

we have

$$\partial u_n(x_n)/\partial\nu_n \leq \partial v_n(x_n)/\partial\nu_n \leq C_0$$

for some $C_0 > 0$, where $\nu_n = (y_n - x_n)/|y_n - x_n|$ and y_n is as in Lemma 3.3.

On the other hand, by Lemma 3.3,

$$\partial u_n(x_n)/\partial\nu_n \geq c_n\partial\psi(x_n)/\partial\nu_n = c_n[2\sigma Re^{-\sigma R^2}] \to \infty$$

as $n \to \infty$. This contradiction finishes the proof. □

LEMMA 3.5. *For any compact set* $K \subset \overline{\Omega}\setminus\overline{\Omega}_0$, $u_n \to U_{a_0}$ *in* $C^{2+\mu}(K)$ *as* $n \to \infty$.

Proof. By Lemma 2.1, $u_n \leq \underline{U}_{a_0}$ on $\Omega\setminus\overline{\Omega}_0$. Since $u_n(x) \leq u_{n+1}(x)$, $\lim_{n\to\infty} u_n(x) = u_\infty(x)$ exists. It follows that u_∞ satisfies (1.3) with $a = a_0$. Here the fact that

$u_\infty = \infty$ on $\partial\Omega_0$ follows from $u_n(x) \le u_{n+1}(x)$ and $u_n(x) \to \infty$ uniformly on $\partial\Omega_0$ by Lemma 3.4. By Theorem 2.4, we necessarily have $u_\infty = \underline{U}_{a_0}$.

Using Sobolev imbedding theorems and interior estimates, we easily see that $u_n \to \underline{U}_{a_0}$ in $C^{2+\mu}(K)$ as $n \to \infty$, for any compact set $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$. □

From Lemmas 3.4 and 3.5, we obtain immediately the following result.

THEOREM 3.6. *Let $a_0 = \lambda_1^D(\Omega_0)$. Then*

(i) $u_a \to \infty$ *uniformly on* $\overline{\Omega}_0$ *as* $a \nearrow a_0$;

(ii) $u_a \to \underline{U}_{a_0}$ *in* $C^{2+\mu}(K)$, *as* $a \nearrow a_0$, *for any compact set* $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$.

Clearly, Theorem 1.2 is a weaker version of Theorem 3.6.

4. Blow-up solutions of problem (1.2).

In this section, we study how the solutions of (1.2) with admissible nontrivial nonnegative initial values blow up as $t \to \infty$ when $a \ge a_0 = \lambda_1^D(\Omega_0)$.

Denote $X = C(\overline{\Omega})$ and $X^+ = \{u \in X : u \ge 0 \ \ \forall x \in \Omega\}$.

LEMMA 4.1. *Suppose $a \ge a_0$ and $u_0 \in X^+ \setminus \{0\}$. Then the unique solution $u(t,x)$ of* (1.2) *satisfies*

$$\lim_{t\to\infty} u(t,x) = \infty \qquad \textit{uniformly for } \ x \in \overline{\Omega}_0.$$

Proof. For $\epsilon > 0$ small, let $u_\epsilon(t,x)$ be the unique solution to the problem

$$\left\{\begin{array}{ll} u_t - \Delta u = (a_0-\epsilon)u - b(x)u^p, & (t,x) \in (0,\infty)\times\Omega, \\ Bu = 0, & (t,x) \in (0,\infty)\times\partial\Omega, \\ u(0,x) = u_0(x), & x \in \Omega. \end{array}\right. \tag{4.1}$$

Since $a > a_0 - \epsilon$, clearly, $u(t,x)$ is a supersolution of (4.1) and hence

$$u(t,x) \ge u_\epsilon(t,x) \ \ \forall (t,x) \in [0,\infty)\times\Omega.$$

For any given $M > 0$, by Lemma 3.4, we can find $\epsilon_0 > 0$ such that the unique positive solution $u_{a_0-\epsilon_0}(x)$ of

$$-\Delta u = (a_0-\epsilon_0)u - b(x)u^p \ \text{ in } \Omega, \quad Bu|_{\partial\Omega} = 0$$

satisfies $u_{a_0-\epsilon_0}(x) > M \ \forall\, x \in \overline{\Omega}_0$. But it is well known that $u_{\epsilon_0}(t,x) \to u_{a_0-\epsilon_0}(x)$ as $t \to \infty$ in the $L^\infty(\Omega)$ norm. Hence $\min_{x\in\overline{\Omega}_0} u(t,x) \ge M \ \forall$ large t. This implies that $u(t,x) \to \infty$ as $t \to \infty$ uniformly for $x \in \overline{\Omega}_0$. □

LEMMA 4.2. *Suppose $a \ge a_0$ and $u_0 \in X^+ \setminus \{0\}$. Then the unique solution $u(t,x)$ of* (1.2) *satisfies*

(i) $\underline{\lim}_{t\to\infty} u(t,x) \ge \underline{U}_a(x)$ *and* $\overline{\lim}_{t\to\infty} u(t,x) \le \overline{U}_a(x) \ \ \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0$;

(ii) *if* (1.3) *has a unique positive solution denoted as U_a, then* $\lim_{t\to\infty} u(t,\cdot) = U_a$ *in $C^{2+\mu}(K)$ for any compact set $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$.*

Proof. Let us denote the unique positive solution of (2.3) with $\phi \equiv c > 0$ by u_c. We first show that for any $v_0 \in C(\overline{\Omega} \setminus \Omega_0)$ with $v_0 \ge 0$, $v_0 \not\equiv 0$, the unique solution $v(t,x)$ of

$$\left\{\begin{array}{ll} v_t - \Delta v = av - b(x)v^p, & (t,x) \in (0,\infty)\times[\Omega\setminus\overline{\Omega}_0], \\ v(t,x) = c, & (t,x) \in (0,\infty)\times\partial\Omega_0, \\ Bv(t,x) = 0, & (t,x) \in (0,\infty)\times\partial\Omega, \\ v(0,x) = v_0(x), & x \in \Omega\setminus\overline{\Omega}_0 \end{array}\right. \tag{4.2}$$

satisfies

$$v(t,x) \to u_c(x) \text{ as } t \to \infty \ \text{ uniformly for } x \in \overline{\Omega} \setminus \Omega_0.$$

Indeed, for any constant $M > 1$, Mu_c is a supersolution of (2.3). Choose $M > 1$ large such that $Mu_c(x) > v(1,x)$ on $\Omega \setminus \Omega_0$ (note this is possible by the strong maximum principle), and let $\underline{v}(t,x)$ denote the unique solution of (4.2) with $v(0,x) \equiv 0$, and let $\overline{v}(t,x)$ denote the unique solution of (4.2) with $v(0,x) \equiv Mu_c(x)$. Then $\underline{v}(t+1,x) \le v(t+1,x) \le \overline{v}(t,x)$. But since u_c is the only steady-state solution of (4.2), we have $\underline{v}(t,x) \to u_c(x)$ and $\overline{v}(t,x) \to u_c(x)$ as $t \to \infty$ uniformly for $x \in \overline{\Omega} \setminus \Omega_0$ (see, e.g., [Ma] or [Sa]). Hence $v(t,x) \to u_c(x)$ as $t \to \infty$ uniformly for $x \in \overline{\Omega} \setminus \Omega_0$.

By Lemma 2.1 and the properties of $\underline{U}_a$, for any $c > 0$, $u_c \le \underline{U}_a$ and $c \to u_c(x)$ is increasing. It then follows easily from the equation of u_c that $u_c \to \underline{U}_a$ as $c \to \infty$ in $C^{2+\mu}(K)$ for any compact set $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$.

For any give $\epsilon > 0$, let $c_0 > 0$ be chosen such that

$$u_{c_0}(x) > \underline{U}_a(x) - \epsilon/2 \ \ \forall x \in K. \tag{4.3}$$

By Lemma 4.1 we can find $T > 0$ such that the unique solution of (1.2) with initial value $u_0 \in X^+ \setminus \{0\}$ satisfies $u(t,x) > c_0$ for $t \ge T$ and $x \in \partial\Omega_0$. Therefore, $u(T+t,x)$ is a supersolution of (4.2) with $c = c_0$ and $v_0(x) = u(T,x)$. Hence, $u(t,x) \ge u_{c_0}(x) - \epsilon/2$ for all large t and $x \in \overline{\Omega} \setminus \overline{\Omega}_0$. Using (4.3), we see that $\forall$ large t,

$$u(t,x) \ge \underline{U}_a(x) - \epsilon \ \ \forall x \in K. \tag{4.4}$$

This implies that

$$\underline{\lim}_{t\to\infty} u(t,x) \ge \underline{U}_a(x) \ \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0. \tag{4.5}$$

Choose a large constant $M > 1$ such that $M\underline{U}_a(x) > u(1,x)$ for all $x \in \overline{\Omega} \setminus \overline{\Omega}_0$ (note this is possible in the Dirichlet boundary condition case since $\underline{U}_a(x) \ge u_{a'}(x)$ near $\partial\Omega$, where $a' \in (\lambda_1(\Omega), \lambda_1^D(\Omega_0))$ and $\partial u_{a'}/\partial\nu < 0$ on $\partial\Omega$). An application of the parabolic maximum principle shows that

$$u(t,x) \le M\underline{U}_a(x) \ \ \forall t > 1, x \in \overline{\Omega} \setminus \overline{\Omega}_0. \tag{4.6}$$

Now consider (4.2) with Ω_0 replaced by Ω_n, where $\Omega_n = \{x \in \Omega : d(x,\Omega_0) < 1/n\}$. We may assume that $\overline{\Omega}_n \subset\subset \Omega \ \forall \ n \ge 1$. Then, for any fixed n, as before, every solution $v_{n,c}(t,x)$ of (4.2) (with Ω_0 replaced by Ω_n) converges to the unique steady-state solution, which we denote as $v^*_{n,c}(x)$. By the parabolic maximum principle and (4.6), $\forall$ large c, $u(t,x) \le v_{n,c}(t,x)$ for $(t,x) \in (0,\infty) \times \overline{\Omega} \backslash \Omega_n$, provided $v_0(x) = u(0,x)$. Thus

$$\overline{\lim}_{t\to\infty} u(t,x) \le v^*_{n,c}(x) \ \ \forall x \in \overline{\Omega} \setminus \Omega_n. \tag{4.7}$$

By the proof of Theorem 2.4, as $c \to \infty$, $v^*_{n,c}(x) \to v^*_{n,\infty}(x)$, which is the minimal positive solution of (1.3) with Ω_0 replaced by Ω_n, and as $n \to \infty$, $v^*_{n,\infty}(x) \to \overline{U}_a(x)$, the maximal solution of (1.3). Therefore, by (4.7),

$$\overline{\lim}_{t\to\infty} u(t,x) \le \overline{U}_a(x) \ \ \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0. \tag{4.8}$$

If (1.3) has a unique positive solution U_a, then by (4.5) and (4.8), we necessarily have

$$\lim_{t\to\infty} u(t,x) = U_a(x) \ \ \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0.$$

Moreover, if $\{t_n\}$ is any sequence of positive numbers satisfying $t_n \to \infty$ as $n \to \infty$, then it follows from (4.6) that $\{u_\infty(t_k, .) : k \geq 1\}$ is bounded in $C(K)$. By regularity, it is compact in $C^{2+\mu}(K)$. Hence we can find a subsequence of $\{t_k\}$, still denoted by $\{t_k\}$, and some C^2 function u_∞ such that

$$u_\infty(t_k, .) \to u_\infty \text{ in } C^{2+\mu}(K).$$

But we must have $u_\infty = U_a$ by our previous discussions. This implies that $\lim_{t\to\infty} u(t, .) = U_a$ in the $C^{2+\mu}(K)$ norm. The proof of Lemma 4.2 is now complete. □

Thus we have the following result which implies Theorem 1.3.

Theorem 4.3. *Suppose $a \geq a_0 = \lambda_1^D(\Omega_0)$ and $u_0 \in X^+ \setminus \{0\}$. Then the unique solution $u(t, x)$ of* (1.2) *satisfies*

(i) $\lim_{t\to\infty} u(t, x) = \infty$ *uniformly for* $x \in \overline{\Omega}_0$.

(ii) $\underline{\lim}_{t\to\infty} u(t, x) \geq \underline{U}_a(x)$ *and* $\overline{\lim}_{t\to\infty} u(t, x) \leq \overline{U}_a(x)$ $\forall x \in \overline{\Omega} \setminus \overline{\Omega}_0$;

(iii) *if* (1.3) *has a unique positive solution denoted as U_a, then* $\lim_{t\to\infty} u(t, .) = U_a$ *in $C^{2+\mu}(K)$ for any compact set $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$.*

As we are able to show that (1.3) has a unique positive solution only if certain conditions on $b(x)$ are satisfied (see Theorem 2.8), it is of some interest to see whether conclusion (ii) in Theorem 4.3 can be improved. We suspect that $\lim_{t\to\infty} u(t, x) = \underline{U}_a(x)$, but again, can only prove this under some conditions on $b(x)$.

Proposition 4.4. *Let us denote $d(x) = d(x, \Omega_0)$. If there exist constants $\alpha \geq 0, c_1 > 0$, $c_2 > 0$ and $\hat{\alpha} \in (\alpha, \alpha + 2)$ such that*

$$c_1 d(x)^{\hat{\alpha}} \leq b(x) \leq c_2 d(x)^\alpha \ \forall \ x \text{ near } \ \partial\Omega_0, \tag{4.9}$$

then for $a \geq a_0$ and $u_0 \in X^+ \setminus \{0\}$, the unique solution $u(t, x)$ of (1.2) *satisfies* $\lim_{t\to\infty} u(t, .) = \underline{U}_a$ *in $C^{2+\mu}(K)$ for any compact set $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$.*

Proof. To make the ideas more transparent, we divide the proof into two steps.

Step 1. The conclusion holds under the condition

$$\underline{\lim}_{d(x)\to 0} b(x)[\underline{U}_a(x)]^{p-1} > a/p. \tag{4.10}$$

We first show that there exists some large positive constant C such that

$$u(t, x) - \underline{U}_a(x) \leq C \ \forall (t, x) \in (0, \infty) \times \overline{\Omega} \setminus \overline{\Omega}_0. \tag{4.11}$$

If this is not true, then we can find (t_n, x_n) satisfying

$$u(t_n, x_n) - \underline{U}_a(x_n) = \max\{u(t, x) - \underline{U}_a(x) : 0 \leq t \leq n, \, x \in \overline{\Omega} \setminus \overline{\Omega}_0\} \to \infty,$$

as $n \to \infty$. By (4.6) and the fact that $\underline{U}_a(x)$ blows up at $\partial\Omega_0$, one easily sees that

$$x_n \in \Omega \setminus \overline{\Omega}_0, \quad d(x_n) \to 0. \tag{4.12}$$

We now look at the equation satisfied by $u(t, x) - \underline{U}_a(x)$:

$$\frac{\partial}{\partial t}(u - \underline{U}_a) - \Delta(u - \underline{U}_a) = \Big\{a - pb(x)\big[\theta u^{p-1} + (1 - \theta)\underline{U}_a^{p-1}\big]\Big\}(u - \underline{U}_a), \tag{4.13}$$

where $\theta = \theta(t, x) \in (0, 1)$. At $(t, x) = (t_n, x_n)$, because of (4.12), the left side of (4.13) is nonnegative. However, using (4.10), we see that the right side of (4.13) at $(t, x) = (t_n, x_n)$ is negative! This contradiction proves (4.11).

Given any small $\epsilon > 0$, denote $\Omega_\epsilon = \{x \in \Omega : d(x, \Omega_0) < \epsilon\}$. By (4.11), we know that

$$u(t,x) \le (1+\delta(\epsilon))\underline{U}_a(x) \quad \forall t \ge 0, x \in \partial\Omega_\epsilon, \tag{4.14}$$

where $\delta(\epsilon) > 0$ and converges to 0 as $\epsilon \to 0$. Let $v_\epsilon(t,x)$ denote the unique solution of

$$\begin{cases} v_t - \Delta v = av - b(x)v^p, & (t,x) \in (0,\infty) \times [\Omega \setminus \overline{\Omega}_\epsilon], \\ v(t,x) = (1+\delta(\epsilon))\underline{U}_a(x), & (t,x) \in (0,\infty) \times \partial\Omega_\epsilon, \\ Bv(t,x) = 0, & (t,x) \in (0,\infty) \times \partial\Omega, \\ v(0,x) = M\underline{U}_a(x), & x \in \Omega \setminus \overline{\Omega}_\epsilon. \end{cases} \tag{4.15}$$

Then using (4.6), (4.14), and the parabolic maximum principle, we obtain

$$u(t+1,x) \le v_\epsilon(t,x) \quad \forall t \ge 0, x \in \overline{\Omega} \setminus \overline{\Omega}_\epsilon. \tag{4.16}$$

By Lemma 2.3, problem (4.15) has a unique steady-state solution which we denote as $u_\epsilon(x)$. It follows that $v_\epsilon(t,x) \to u_\epsilon(x)$ as $t \to \infty$. Using Lemma 2.1, we easily deduce $u_\epsilon(x) \le (1+\delta(\epsilon))\underline{U}_a(x)$. Therefore, from (4.16), we have

$$\overline{\lim}_{t\to\infty} u(t,x) \le (1+\delta(\epsilon))\underline{U}_a(x) \quad \forall x \in \overline{\Omega} \setminus \overline{\Omega}_\epsilon.$$

Letting $\epsilon \to 0$, we finally obtain

$$\overline{\lim}_{t\to\infty} u(t,x) \le \underline{U}_a(x) \quad \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0.$$

This combined with conclusion (ii) of Theorem 4.3 gives

$$\lim_{t\to\infty} u(t,x) = \underline{U}_a(x) \quad \forall x \in \overline{\Omega} \setminus \overline{\Omega}_0.$$

As in the proof of Lemma 4.2, it follows then from the regularity of solutions that $\lim_{t\to\infty} u(t,.) = \underline{U}_a$ in $C^{2+\mu}(K)$ for any compact set $K \subset \overline{\Omega} \setminus \overline{\Omega}_0$.

Step 2. Condition (4.9) implies (4.10).

By (4.9), we can find $c_2' > c_2$ and $c_1' \in (0, c_1)$ such that

$$c_1' d(x)^{\hat{\alpha}} \le b(x) \le c_2' d(x)^\alpha \ \forall x \in \overline{\Omega} \setminus \Omega_0.$$

Hence, one easily sees that $\underline{U}_a(x)$ is a supersolution to (1.3) with $b(x)$ replaced by $c_2' d(x)^\alpha$, whose unique positive solution (guaranteed by Theorem 2.8 and Remark 2.9) we denote as $v(x)$. Considering that $v(x)$ can be obtained as the limit of the solutions of problem (2.3) with $\phi \equiv n \to \infty$ and $b(x) = c_2' d(x)^\alpha$, one sees easily by Lemma 2.1 that $\underline{U}_a(x) \ge v(x)$. By Theorem 2.8 and Remark 2.9,

$$\lim_{d(x)\to 0} d(x)^{\alpha+2} v(x)^{p-1} = c_0 > 0.$$

Hence,

$$\underline{\lim}_{d(x)\to 0} b(x)[\underline{U}_a(x)]^{p-1} \ge \lim_{d(x)\to 0} c_1' d(x)^{\hat{\alpha}} v(x)^{p-1} = +\infty > a/p.$$

The proof is complete. □

Finally we give some estimates on the blow-up rate of the solutions of (1.2).

Theorem 4.5. *Let $a \geq a_0$ and $u_0 \in X^+ \setminus \{0\}$. Then for any given small $\epsilon > 0$, there exists a constant $M_\epsilon > 0$ such that the unique solution $u(t,x)$ of* (1.2) *satisfies*

$$u(t,x) \leq M_\epsilon e^{(a-a_0+\epsilon)t} \quad \forall \quad x \in \overline{\Omega}_0 \quad \text{and } \forall \text{ large } t;$$

for any $x \in \Omega_0$ and any positive constant M, it holds

$$u(t,x) \geq M e^{(a-a_0)t} \quad \forall \text{ large } t.$$

Proof. For any given $\epsilon > 0$, define

$$v(t,x) = e^{(a_0-\epsilon-a)t}u(t,x),$$

where $u(t,x)$ is the unique solution of (1.2). A simple calculation shows

$$v_t - \Delta v \leq (a_0 - \epsilon)v - b(x)v^p.$$

Clearly, $Bv|_{\partial\Omega} = 0$ and $v(0,x) = u(0,x)$. Hence v is a subsolution of (1.2) with a replaced by $a_0 - \epsilon$. It follows from the parabolic maximum principle that

$$v(t,x) \leq u_\epsilon(t,x) \quad \forall (t,x) \in (0,\infty) \times \Omega.$$

Here $u_\epsilon(t,x)$ denotes the positive solution of (1.2) with a replaced by $a_0 - \epsilon$. It is well known that $u_\epsilon(t,x) \to u_{a_0-\epsilon}(x)$ uniformly on $\overline{\Omega}$. Hence if we define M_ϵ by

$$M_\epsilon = (1+\epsilon) \max_{x \in \overline{\Omega}} u_{a_0-\epsilon}(x),$$

then $v(t,x) \leq M_\epsilon$ $\forall$ large t and $x \in \overline{\Omega}_0$. That is,

$$u(t,x) \leq M_\epsilon e^{(a-a_0+\epsilon)t} \quad \forall \quad x \in \overline{\Omega}_0 \quad \text{and } \forall \text{ large } t.$$

Next we let $\phi_0(x) \geq 0$ with $\|\phi_0\|_{L^\infty(\Omega_0)} = 1$ be the eigenfunction for the problem

$$-\Delta\phi = a_0\phi, \quad \phi|_{\partial\Omega_0} = 0.$$

Then for any positive constant M', $v(t,x) = M'e^{(a-a_0)t}\phi_0(x)$ satisfies

$$v_t - \Delta v = av \text{ in } \Omega_0, \quad v|_{\partial\Omega_0} = 0.$$

By Lemma 4.1, there exists $T > 0$ such that $u(T,x) > M'\phi_0(x)$ on $\overline{\Omega}_0$. Hence by the parabolic maximim principle, $v(t,x) \leq u(T+t,x)$ for all $t > 0$ and $x \in \Omega_0$. Now for any given $x \in \Omega_0$ and $M > 0$, choose M' such that $M'e^{-(a-a_0)T}\phi_0(x) = M$; then

$$u(t,x) \geq M e^{(a-a_0)t} \quad \forall \text{ large } t.$$

This finishes the proof of Theorem 4.5. □

Acknowledgments. We thank Professor E. N. Dancer for useful comments on this paper. We also thank the referees for their careful reading of the paper and suggestions.

REFERENCES

[AT] S. Alama and G. Tarantello, *On the solvability of a semilinear elliptic equation via an associated eigenvalue problem*, Math. Z., 221 (1996), pp. 467–493.

[AT2] S. Alama and G. Tarantello, *On semilinear elliptic equations with indefinite nonlinearities*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 439–475.

[AG] A. Ambrosetti and J. L. Gámez, *Branches of positive solutions for some semilinear Schrödinger equations*, Math. Z., 224 (1997), pp. 347–362.

[BM] C. Bandle and M. Marcus, *"Large" solutions of semilinear elliptic equations: Existence, uniqueness and asymptotic behavior*, J. Anal. Math., 58 (1992), pp. 9–24.

[BBL] R. Benguria, H. Brézis, and E. Lieb, *The Thomas–Fermi–Von Weizsacker theory of atoms and molecules*, Comm. Math. Phys., 79 (1981), pp. 167–180.

[BCN] H. Berestycki, I. Capuzzo-Dolcetta, and L. Nirenberg, *Variational methods for indefinite superlinear homogeneous elliptic problems*, in Non Linear Differential Equations and Applications, to appear.

[BNV] H. Beresticky, L. Nirenberg, and S. R. S. Varadhan, *The principal eigenvalue and maximum principle for second-order elliptic operators in general domains*, Comm. Pure Appl. Math., 47 (1994), pp. 47–92.

[Da] E. N. Dancer, *Some remarks on classical problems and fine properties of Sobolev spaces*, Differential Integral Equations, 9 (1996), pp. 437–446.

[dP] M. A. del Pino, *Positive solutions of a semilinear elliptic equation on a compact manifold*, Nonlinear Anal., 22 (1994), pp. 1423–1430.

[DG] Y. Du and Z. M. Guo, *The degenerate logistic model and a singularly mixed boundary blow-up problem*, submitted.

[FKLM] J. M. Fraile, P. Koch Medina, J. López-Gómez, and S. Merino, *Elliptic eigenvalue problems and unbounded continua of positive solutions of a semilinear elliptic equation*, J. Differential Equations, 127 (1996), pp. 295–319.

[Fr] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[GGLS] J. Gárcia-Melián, R. Gómez-Reñasco, J. López-Gómez, and J. Sabina de Lis, *Point-wise growth and uniqueness of positive solutions for a class of sublinear elliptic problems where bifurcation from infinity occurs*, Arch. Rational Mech. Anal., 145 (1998), pp. 261–289.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, New York, 1977.

[Hs] P. Hess, *Periodic-Parabolic Boundary Value Problems and Positivity*, Longman Scientific and Technical, Harlow, UK, 1991.

[KW] J. L. Kazdan and F. W. Warner, *Scalar curvature and conformal deformation of Riemannian structure*, J. Differential Geometry, 10 (1975), pp. 113–134.

[LSU] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uraceva, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc., Providence, RI, 1967.

[LM] A. C. Lazer and P. J. McKenna, *On a problem of Bieberbach and Redemacher*, Nonlinear Anal., 21 (1993), pp. 327–335.

[LN] C. Loewner and L. Nirenberg, *Partial differential equations invariant under conformal or projective transformations*, in Contributions to Analysis, L. V. Ahlfors et al., eds., Academic Press, New York, 1974, pp. 245–272.

[LS] J. López-Gómez and J. C. Sabina de Lis, *First variations of principal eigenvalues with respect to the domain and point-wise growth of positive solutions for problems where bifurcation from infinity occurs*, J. Differential Equations, 148 (1998), pp. 47–64.

[MV] M. Marcus and L. Véron, *Uniqueness and asymptotic behavior of solutions with boundary blow-up for a class of nonlinear elliptic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 237–274.

[Ma] H. Matano, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. Res. Inst. Math. Sci., 15 (1979), pp. 401–454.

[Ou] T. Ouyang, *On the positive solutions of semilinear equations* $\Delta u + \lambda u - hu^p = 0$ *on the compact manifolds*, Trans. Amer. Math. Soc., 331 (1992), pp. 503–527.

[Sa] D. H. Sattinger, *Monotone methods in nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 21 (1972), pp. 979–1000.

[St] K. Schmitt, *Boundary value problems for quasilinear second-order elliptic equations*, Nonlinear Anal., 2 (1978), pp. 263–309.

SIAM J. MATH. ANAL.
Vol. 31, No. 1, pp. 19–33

A SOBOLEV SPACE THEORY OF SPDEs WITH CONSTANT COEFFICIENTS IN A HALF SPACE[*]

N. V. KRYLOV[†] AND S. V. LOTOTSKY[‡]

Abstract. Equations of the form $du = (a^{ij}u_{x^ix^j} + D_if^i)\,dt + \sum_k(\sigma^{ik}u_{x^i} + g^k)\,dw_t^k$ are considered for $t > 0$ and $x \in \mathbb{R}^d_+$. The unique solvability of these equations is proved in weighted Sobolev spaces with fractional positive or negative derivatives, summable to the power $p \in [2,\infty)$.

Key words. stochastic partial differential equations, Sobolev spaces with weights

AMS subject classifications. 60H15, 35R60

PII. S0036141098338843

Introduction. The main goal of this article is to extend the results of [6] to multidimensional cases. We are dealing with the equation

$$du = (a^{ij}u_{x^ix^j} + f^i_{x^i})\,dt + (\sigma^{ik}u_{x^i} + g^k)\,dw_t^k$$

given for $t \geq 0$ and $x \in \mathbb{R}^d_+ := \{x = (x^1, x') : x^1 > 0, x' \in \mathbb{R}^{d-1}\}$. Here w_t^k are independent one-dimensional Wiener processes, i and j run from 1 to d, k runs through $\{1, 2, \dots\}$ with the summation convention being enforced, and f^i and g^k are some given functions of (ω, t, x) defined for $i = 1, \dots, d$ and $k \geq 1$. The functions a^{ij} and σ^{ik} are assumed to depend only on ω and t, and in this sense we consider equations with "constant" coefficients. Without loss of generality we also assume that $a^{ij} = a^{ji}$.

As in [6], let us mention that such equations with a finite number of the processes w_t^k appear, for instance, in nonlinear filtering problems for partially observable diffusions (see [8]). Considering infinitely many w_t^k turns out to be instrumental in treating equations for measure valued processes, for instance, driven by space-time white noise (see [3] or [4]).

Our main goal is to prove the solvability of such equations in spaces similar to Sobolev spaces, in which derivatives are understood as generalized functions, the number of derivatives may be fractional or negative, and underlying power of summability is $p \in [2,\infty)$.

The motivation for this goal is explained in detail in [3] or [4], where an L_p-theory is developed for the equations in the whole space. We mention only that if $p = 2$, the theory was developed long ago and an account of it can be found, for instance, in [8]. The case of equations in domains is also treated in [8]. However, the solvability is proved only in spaces W_2^1 of functions having one generalized derivative in x square summable in (ω, t, x). It turns out that going to better smoothness of solutions is not possible in spaces W_2^n and one needs to consider Sobolev spaces with weights, allowing derivatives to blow up near the boundary. The theory of solvability in Hilbert spaces like W_2^n with weights is developed in [1] and [7], where n is an integer. Here we show

[*] Received by the editors May 18, 1998; accepted for publication (in revised form) April 9, 1999; published electronically November 4, 1999.

http://www.siam.org/journals/sima/31-1/33884.html

[†] 127 Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (krylov@math.umn.edu). The work of this author was partially supported by NSF grant DMS–9625483.

[‡] Department of Mathematics, M.I.T., Room 2-247, 77 Massachusetts Avenue, Cambridge, MA 02139-4307 (lototsky@math.mit.edu).

what happens if one takes a fractional or negative number of derivatives and replaces 2 with any $p \geq 2$. By the way, according to [2], it is not possible to take $p < 2$ when a stochastic term is present in the equation.

One of the main difficulties in developing the theory presented below was finding the right spaces where to look for solutions. In the one-dimensional case $\mathbb{R}^d_+ = \mathbb{R}_+$ they have been found in [6]. It turns out that there are many multidimensional counterparts of spaces from [6]. The one which looks the most natural is to apply weights only to derivatives with respect to x^1. Indeed, why should we allow the derivatives with respect to tangential variables blow up near $x^1 = 0$? The equation is translation invariant with respect to x', isn't it? However, in such spaces it is impossible to solve equations with variable coefficients in smooth domains unless the coefficients not only are smooth with respect to x but also behave in a very restrictive way as x approaches the boundary. And, of course, considering equations with constant coefficients in half spaces aims at equations with variable coefficients in smooth domains.

This shows that one cannot just imitate the original definition of Sobolev spaces with weights $H^\gamma_{p,\theta}$ from [6]. However, it turns out that one can very naturally generalize to the multidimensional case an equivalent definition, looking much more complex, which is discovered in [6] and stated there as Theorem 1.11 (see Definition 1.1 below).

This article is organized as follows. In section 1 we present some definitions and facts from [5] on the basis of which, in section 2, we introduce the stochastic Banach spaces in which we are going to solve our equations. Our main result is given and proved in section 3. One auxiliary result used in section 3 is proved in section 4.

We finish the introduction with some notation. Everywhere, apart from section 1, we assume that $p \in [2, \infty)$. By $C^n_0(D)$ we denote the set of all n times continuously differentiable (real-valued) functions with compact support belonging to D. We denote

$$D_i = \partial/\partial x^i, \quad Du = u_x = (D_1 u, \dots, D_d u).$$

For a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$, where α_i's are nonnegative integers, denote

$$D^\alpha = D_1^{\alpha_1} \cdots D_d^{\alpha_d}, \quad |\alpha| = \alpha_1 + \dots + \alpha_d.$$

By $H^\gamma_p = H^\gamma_p(\mathbb{R}^d)$ we denote the space of Bessel potentials ($= (1-\Delta)^{-\gamma/2} L_p$) with norm $|| \cdot ||_{\gamma,p}$ (see [9]). For $\gamma = 0$, we have $H^0_p = L_p$ and we denote $|| \cdot ||_p = || \cdot ||_{0,p}$.

Any function given on $\mathbb{R}_+ := \mathbb{R}^1_+$ is also considered as a function on $\mathbb{R}^d_+$ independent of x'. Define M^α as the operator of multiplying by $(x^1)^\alpha$, $M = M^1$.

Finally, by $\mathcal{D}(R^d_+)$ we denote the space of all distributions on $\mathbb{R}^d_+$ that is of continuous linear functionals on $C^\infty_0(\mathbb{R}^d_+)$.

1. Sobolev spaces with weights.

Here we collect some definitions and facts from [5].

DEFINITION 1.1. *Take and fix a nonnegative function $\zeta \in C^\infty_0(\mathbb{R}_+)$ such that*

$$\sum_{n=-\infty}^{\infty} \zeta^p(e^{x-n}) \geq 1 \quad \textit{for all} \;\; x \in \mathbb{R}. \tag{1.1}$$

For $\gamma, \theta \in \mathbb{R}$, and $p \in (1, \infty)$ let $H^\gamma_{p,\theta}$ be the set of all distributions u on $\mathbb{R}^d_+$ such that

$$||u||^p_{\gamma,p,\theta} := \sum_{n=-\infty}^{\infty} e^{n\theta} ||u(e^n \cdot)\zeta||^p_{\gamma,p} = \sum_{n=-\infty}^{\infty} e^{n\theta} ||(1-\Delta)^{\gamma/2}(u(e^n \cdot)\zeta)||^p_p < \infty. \tag{1.2}$$

Denote $L_{p,\theta} = H^0_{p,\theta}$.

In the same way, for any separable Banach space X, *we introduce the spaces* $H^\gamma_{p,\theta}(X)$ *of* X*-valued functions by replacing* $(1-\Delta)^{\gamma/2}(u(e^n \cdot)\zeta)$ *in* (1.2) *with* $|(1-\Delta)^{\gamma/2}(u(e^n \cdot)\zeta)|_X$.

LEMMA 1.2. (i) *The spaces* $H^\gamma_{p,\theta}$ *are Banach spaces and the space* $C_0^\infty(\mathbb{R}^d_+)$ *is dense in* $H^\gamma_{p,\theta}$.

(ii) *For different* ζ *satisfying* (1.1), *we get the same spaces* $H^\gamma_{p,\theta}$ *with equivalent norms. Furthermore, if* $\eta \in C_0^\infty(\mathbb{R}^d_+)$, *then for any* $u \in \mathcal{D}(\mathbb{R}^d_+)$ *and* γ, θ, p *we have*

$$\sum_{n=-\infty}^{\infty} e^{n\theta} ||u(e^n \cdot)\eta||^p_{\gamma,p} \leq N \sum_{n=-\infty}^{\infty} e^{n\theta} ||u(e^n \cdot)\zeta||^p_{\gamma,p},$$

where N *depends only on* γ, θ, p, η, d *(and* ζ*).*

(iii) *Let* $\alpha \in \mathbb{R}$. *We have* $u \in H^\gamma_{p,\theta}$ *if and only if* $u = M^\alpha v$ *with* $v \in H^\gamma_{p,\theta+\alpha p}$. *Hence,*

$$M^\alpha H^\gamma_{p,\theta+\alpha p} = H^\gamma_{p,\theta}.$$

In addition,

$$||u||_{\gamma,p,\theta} \leq N||M^{-\alpha}u||_{\gamma,p,\theta+\alpha p} \leq N||u||_{\gamma,p,\theta},$$

where N *are independent of* u.

(iv) *The space* $L_{p,\theta}$ *coincides with the space of functions summable to the power* p *over* $\mathbb{R}^d_+$ *with respect to the measure* $(x^1)^{\theta-d}\,dx$.

(v) *If* γ *is a nonnegative integer, then the space* $H^\gamma_{p,\theta}$ *is*

$$\{u : u, x^1 u_x, \dots, (x^1)^{|\alpha|} D^\alpha u \in L_{p,\theta} \quad \textit{for all} \ \ \alpha : |\alpha| \leq \gamma\}$$

with a natural norm.

The spaces $H^\gamma_{p,\theta}$ are introduced and studied in [5] for all $\theta \in \mathbb{R}$. However, below in this article we always suppose that $d-1 < \theta < p+d-1$. For this range of θ, the following results, again borrowed from [5], are true.

LEMMA 1.3. *Let* $d-1 < \theta < p+d-1$.

(i) *The following conditions are equivalent:*

(a) $u \in H^\gamma_{p,\theta}$,

(b) $u \in H^{\gamma-1}_{p,\theta}$ *and* $Mu_x \in H^{\gamma-1}_{p,\theta}$,

(c) $u \in H^{\gamma-1}_{p,\theta}$ *and* $(Mu)_x \in H^{\gamma-1}_{p,\theta}$.

In addition, under either of these three conditions for some constants $N = N(\gamma, p, \theta, d)$ *we have*

$$||u||_{\gamma,p,\theta} \leq N||Mu_x||_{\gamma-1,p,\theta} \leq N||u||_{\gamma,p,\theta},$$

$$||u||_{\gamma,p,\theta} \leq N||(Mu)_x||_{\gamma-1,p,\theta} \leq N||u||_{\gamma,p,\theta}.$$

(ii) *We have* $M^{-1}u \in H^\gamma_{p,\theta}$ *if and only if*

$$u_x \in H^{\gamma-1}_{p,\theta} \quad \textit{and} \quad M^{-1}u \in \bigcup_\mu H^\mu_{p,\theta}.$$

Moreover, there exist constants $N = N(d, \gamma, \mu, \theta, p)$ such that, for any $\mu \le \gamma$ and $M^{-1}u \in H^{\gamma}_{p,\theta}$, we have

$$||M^{-1}u||_{\gamma,p,\theta} \le N||u_x||_{\gamma-1,p,\theta} \le N||M^{-1}u||_{\gamma,p,\theta}.$$

(iii) *The operator $\mathcal{L} := M^2\Delta + 2MD_1$ is a bounded operator from $H^{\gamma}_{p,\theta}$ onto $H^{\gamma-2}_{p,\theta}$ and its inverse is also bounded.*

(iv) *There is a bounded operator*

$$Q : u \in H^{\gamma}_{p,\theta} \to Qu = (Q_1 u, \dots, Q_d u) \in (H^{\gamma+1}_{p,\theta})^d$$

such that, for any $u \in H^{\gamma}_{p,\theta}$, we have $u = MD_iQ_iu$.

2. Stochastic Banach spaces on $\mathbb{R}^d_+$. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, $(\mathcal{F}_t, t \ge 0)$ be an increasing filtration of σ-fields $\mathcal{F}_t \subset \mathcal{F}$ containing all P-null subsets of Ω, and $\mathcal{P}$ be the predictable σ-field generated by $(\mathcal{F}_t, t \ge 0)$. Let $\{w^k_t; k = 1, 2, \dots\}$ be a family of independent one-dimensional $\mathcal{F}_t$-adapted Wiener processes defined on $(\Omega, \mathcal{F}, P)$. We are going to use the Banach spaces $\mathbb{H}^{\gamma}_p(\tau)$, $\mathbb{H}^{\gamma}_p(\tau, l_2)$, and $\mathcal{H}^{\gamma}_p(\tau)$ introduced in [3] or [4].

Throughout the remaining part of the paper we assume that

$$d - 1 < \theta < p + d - 1.$$

DEFINITION 2.1. *Let τ be a stopping time and f and g^k, $k = 1, 2, \dots$, be $\mathcal{D}(\mathbb{R}^d_+)$-valued $\mathcal{P}$-measurable functions defined on $(\!(0, \tau]\!]$. We write $f \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ and $g \in \mathbb{H}^{\gamma}_{p,\theta}(\tau, l_2)$ if and only if $f \in L_p((\!(0, \tau]\!]; H^{\gamma}_{p,\theta})$ and $g \in L_p((\!(0, \tau]\!]; H^{\gamma}_{p,\theta}(l_2))$, respectively. We also denote*

$$\mathbb{H}^{\gamma}_{p,\theta} = \mathbb{H}^{\gamma}_{p,\theta}(\infty), \quad \mathbb{H}^{\gamma}_{p,\theta}(l_2) = \mathbb{H}^{\gamma}_{p,\theta}(\infty, l_2), \quad \mathbb{L}_{\dots}\dots = \mathbb{H}^0_{\dots}\dots.$$

In the case $f \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ and $g \in \mathbb{H}^{\gamma+1}_{p,\theta}(\tau, l_2)$ we write $(f, g) \in \mathcal{F}^{\gamma}_{p,\theta}(\tau)$ and define

$$||f||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} = E\int_0^{\tau} ||f(t)||^p_{\gamma,p,\theta}\,dt, \quad ||g||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau,l_2)} = E\int_0^{\tau} ||g(t)||^p_{H^{\gamma}_{p,\theta}(l_2)}\,dt,$$

$$||(f,g)||_{\mathcal{F}^{\gamma}_{p,\theta}(\tau)} = ||f||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} + ||g||_{\mathbb{H}^{\gamma+1}_{p,\theta}(\tau,l_2)}.$$

Finally, we introduce spaces of initial data. We write $u_0 \in U^{\gamma}_{p,\theta}$ if and only if $M^{2/p-1}u(0,\cdot) \in L_p(\Omega, \mathcal{F}_0, H^{\gamma-2/p}_{p,\theta})$ (or by Lemma 1.2, part (iii)*, if and only if $u(0,\cdot) \in L_p(\Omega, \mathcal{F}_0, H^{\gamma-2/p}_{p,\theta+2-p})$) and denote*

$$||u(0,\cdot)||^p_{U^{\gamma}_{p,\theta}} = E||M^{2/p-1}u(0,\cdot)||^p_{\gamma-2/p,p,\theta}.$$

DEFINITION 2.2. *For a $\mathcal{D}(\mathbb{R}^d_+)$-valued function u defined on $\Omega \times ([0,\tau] \cap [0,\infty))$ with $u(0,\cdot) \in U^{\gamma}_{p,\theta}$, we write $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ if and only if $M^{-1}u \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ and there exists $(f, g) \in \mathcal{F}^{\gamma-2}_{p,\theta}(\tau)$ such that, for any $\phi \in C^{\infty}_0(\mathbb{R}^d_+)$, with probability one, we have*

$$(u(t,\cdot), \phi) = (u(0,\cdot), \phi) + \int_0^t (M^{-1}f(s,\cdot), \phi)\,ds + \sum_{k=1}^{\infty}\int_0^t (g^k(s,\cdot), \phi)\,dw^k_s \tag{2.1}$$

for all $t \in [0,\tau] \cap [0,\infty)$. *In this situation we also write* $M^{-1}f = \tilde{\mathbb{D}}u$, $g = \tilde{\mathbb{S}}u$,

$$du = M^{-1}f\,dt + g^k\,dw_t^k$$

and we define $\mathfrak{H}^\gamma_{p,\theta,0}(\tau) = \mathfrak{H}^\gamma_{p,\theta}(\tau) \cap \{u: \ u(0,\cdot) = 0\}$,

$$||u||^p_{\mathfrak{H}^\gamma_{p,\theta}(\tau)} = ||u_x||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)} + ||(f,g)||^p_{\mathcal{F}^{\gamma-2}_{p,\theta}(\tau)} + ||u(0,\cdot)||^p_{U^\gamma_{p,\theta}}. \tag{2.2}$$

As always, we drop τ *in* $\mathfrak{H}^\gamma_{p,\theta}(\tau)$ *and* $\mathcal{F}^\gamma_{p,\theta}(\tau)$ *if* $\tau = \infty$.

Remark 2.3. If $u \in \mathfrak{H}^\gamma_{p,\theta}(\tau)$ and $\phi(x) = \phi(x^1)$ with $\phi \in C_0^\infty(\mathbb{R}_+)$, then ϕu lies in $\mathcal{H}^\gamma_p(\tau)$. By Theorem 2.7 of [4] this implies that if $u \in \mathfrak{H}^\gamma_{p,\theta}(\tau)$ and $||u||_{\mathfrak{H}^\gamma_{p,\theta}(\tau)} = 0$, then u is indistinguishable from zero.

Of course, we identify elements of $\mathfrak{H}^\gamma_{p,\theta}(\tau)$ which are indistinguishable.

Remark 2.4 (cf. Remark 2.3 in [4]). Given $u \in \mathfrak{H}^\gamma_{p,\theta}(\tau)$, there exists only one couple of functions f and g in Definition 2.2. Therefore, the notations $M^{-1}f = \tilde{\mathbb{D}}u$, $g = \tilde{\mathbb{S}}u$, and (2.2) make sense.

It is also worth noting that the last series in (2.1) converges uniformly in t on each interval $[0, \tau \wedge T]$, $T \in (0,\infty)$, in probability.

Remark 2.5. It follows from Lemma 1.3 part (ii) that the condition $M^{-1}u \in \mathbb{H}^\gamma_{p,\theta}(\tau)$ can be replaced with

$$M^{-1}u \in \bigcup_\mu \bigcap_{T>0} \mathbb{H}^\mu_{p,\theta}(\tau \wedge T) \quad \text{and} \quad u_x \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau).$$

Also in (2.2), replacing the norm $||u_x||_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)}$ with $||M^{-1}u||_{\mathbb{H}^\gamma_{p,\theta}(\tau)}$ leads to an equivalent norm.

Remark 2.6. In the same way as in Remark 2.6 of [6] one proves that the spaces $\mathfrak{H}^\gamma_{p,\theta}(\tau)$ and $\mathfrak{H}^\gamma_{p,\theta,0}(\tau)$ are Banach spaces.

Remark 2.7. The term $M^{-1}f$ in (2.1) can be replaced with D_if^i for $f^i := Q_if \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$, $i = 1,\dots,d$ (see Lemma 1.3), and the norm $||f||_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)}$ (participating in (2.2)) with $\sum_i ||f^i||_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)}$, the latter leading to an equivalent norm.

Remark 2.8. If $u \in \mathfrak{H}^\gamma_{p,\theta}(\tau)$, then $MD_iu \in \mathfrak{H}^{\gamma-1}_{p,\theta}(\tau)$ for $i = 1,\dots d$, and

$$||MDu||_{\mathfrak{H}^{\gamma-1}_{p,\theta}(\tau)} \le N(\gamma,\theta,p,d)||u||_{\mathfrak{H}^\gamma_{p,\theta}(\tau)}.$$

Indeed, by Lemma 1.3, $M^{-1}(MD_iu) = D_iu \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$ and by Remark 2.7, $du = D_jf^j\,dt + g^k\,dw_t^k$ with $f^j \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$ and $g \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau, l_2)$, so that

$$d(MD_iu) = M^{-1}M^2D_iD_jf^j\,dt + MD_ig^k\,dw_t^k,$$

where $M^2D_iD_jf^j = MD_iMD_jf^j - \delta^{1i}MD_jf^j$. By Lemma 1.3

$$||M^2D_iD_jf^j||_{\mathbb{H}^{\gamma-3}_{p,\theta}(\tau)} \le N\sum_j ||f^j||_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)} \le N||f||_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)},$$

$$||MD_ig||_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau,l_2)} \le N||g||_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau,l_2)},$$

$$\begin{aligned}&||M^{2/p-1}MD_iu(0,\cdot)||_{\gamma-1-2/p,p,\theta}\\ &\quad = ||MD_i(M^{2/p-1}u(0,\cdot)) - \delta^{1i}(2/p-1)M^{2/p-1}u(0,\cdot)||_{\gamma-1-2/p,p,\theta}\\ &\quad \le N||M^{2/p-1}u(0,\cdot)||_{\gamma-2/p,p,\theta}.\end{aligned}$$

THEOREM 2.9. *For any nonnegative integer $n \geq \gamma$, the set*

$$\mathfrak{H}^n_{p,\theta}(\tau) \bigcap \bigcup_{k=1}^{\infty} \bigcap_{T \in (0,\infty)} L_p(\Omega, C([0, \tau \wedge T], C^n_0(G_k))), \tag{2.3}$$

where $G_k = (1/k, k) \times \{|x'| < k\}$, is everywhere dense in $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$.

Proof. Corollary 1.20 of [5] states that there exists a sequence of functions $\eta_k \in C^{\infty}_0(\mathbb{R}_+)$ vanishing near zero and infinity and such that, for any $u \in H^{\gamma}_{p,\theta}$, we have

$$||\eta_k u||_{\gamma,p,\theta} \leq N||u||_{\gamma,p,\theta}, \quad ||\eta_k u - u||_{\gamma,p,\theta} \to 0$$

as $k \to \infty$, where N is independent of k and u. Obviously, if $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, then $\eta_k u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ and by Remark 2.5 and the above result of [5] we get that $\eta_k u \to u$ in $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$.

To prove the theorem it remains to show only that any $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, vanishing outside some G_k, can be approximated by elements of set (2.3). To do this, notice that for such u its $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$-norm is equivalent to $\mathcal{H}^{\gamma}_p(\tau)$-norm. Next, take a function $\xi \in C^{\infty}_0(\mathbb{R}^d_+)$ with unit integral and for $\varepsilon > 0$ define $\xi_{\varepsilon}(x) = \varepsilon^{-d}\xi(x/\varepsilon)$, $u^{(\varepsilon)}(t,x) := \xi_{\varepsilon}(x) * u(t,x)$. It is easy to check that for ε small enough (for instance, such that $u^{(\varepsilon)}$ vanishes when x^1 is close to zero or infinity), we have $u^{(\varepsilon)} \in \mathfrak{H}^n_{p,\theta}(\tau)$ and $u^{(\varepsilon)} \in \mathcal{H}^n_p(\tau)$ for all n. In addition, by well-known properties of mollified functions, $u^{(\varepsilon)}$ converge to u in $\mathcal{H}^{\gamma}_p(\tau)$- and $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$-norm as $\varepsilon \downarrow 0$. Of course, $u^{(\varepsilon)}(t,x)$ is infinitely differentiable with respect to x.

Finally, since $u \in \mathcal{H}^{\gamma}_p(\tau)$, by Theorems 7.1 and 7.2 of [4] we have

$$u \in L_p(\Omega, C([0, \tau \wedge T], H^{\gamma-1}_p)). \tag{2.4}$$

In addition, by Sobolev's embedding theorems and by properties of mollifiers, for any $v \in H^{\gamma-1}_p$ and multi-index α with $|\alpha| = n$,

$$|D^{\alpha} v^{(\varepsilon)}| \leq N||v^{(\varepsilon)}||_{d+n,p} \leq N\varepsilon^{-\kappa}||v||_{\gamma-1,p},$$

where N and κ are independent of v (and ε). This and (2.4) show that

$$u^{(\varepsilon)} \in L_p(\Omega, C([0, \tau \wedge T], C^n_0(\mathbb{R}^d_+))).$$

The theorem is proved.

By repeating the proof of Theorem 2.9 with obvious changes we obtain one more useful result.

THEOREM 2.10. *The statement of Theorem 2.9 remains true if we replace $\mathfrak{H}^n_{p,\theta}(\tau)$ and $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ with $\mathbb{H}^n_{p,\theta}(\tau)$ and $\mathbb{H}^{\gamma}_{p,\theta}(\tau)$, respectively, or with $\mathbb{H}^n_{p,\theta}(\tau, l_2)$ and $\mathbb{H}^{\gamma}_{p,\theta}(\tau, l_2)$, respectively.*

As in the one-dimensional case (cf. [6]), the following embedding theorem presents certain interest.

THEOREM 2.11. *Let $T \in (0,\infty)$ be a constant and let $\tau \leq T$. Then for any function $u \in \mathfrak{H}^{\gamma}_{p,\theta,0}(\tau)$, we have*

$$E \sup_{t \leq \tau} ||u(t,\cdot)||^p_{\gamma-1,p,\theta} \leq N(p,d,\theta,\gamma) T^{(p-2)/2} ||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}. \tag{2.5}$$

To prove this theorem we use the following fact which is similar to Remark 2.2 of [3] or Remark 4.11 of [4]. Its proof can be obtained just by repeating the proof of Lemma 2.12 of [6] and is omitted.

LEMMA 2.12. *Let $T \in (0,\infty)$ be a constant and let $\tau \leq T$. Let $u \in \mathcal{H}^{\gamma}_{p,0}(\tau)$ and $du = f\,dt + g^k\,dw^k_t$. Then for any constant $c > 0$,*

$$E\sup_{t\leq\tau}||u_x(t,\cdot)||^p_{\gamma-2,p} \leq N(p,d)T^{(p-2)/2}(c||u_{xx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} + c^{-1}||f||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} + ||g_x||^p_{\mathbb{H}^{\gamma-2}_p(\tau,l_2)}).$$

Proof of Theorem 2.11. We proceed as in the proof of Theorem 2.11 of [6]. We have

$$E\sup_{t\leq\tau}||u(t,\cdot)||^p_{\gamma-1,p,\theta} \leq \sum_{n=-\infty}^{\infty} e^{n\theta}E\sup_{t\leq\tau}||u(t,e^n\,\cdot)\zeta||^p_{\gamma-1,p}. \tag{2.6}$$

Define $u_n(t,x) := \zeta(x)u(t,e^nx)$ and notice that, since the support of $\zeta(x)u(t,e^nx)$ is not larger than the one of $\zeta(x)$, we have (see, for instance, Remark 1.12 of [5])

$$||u_n(t,\cdot)||_{\gamma-1,p} \leq N||u_{nx}(t,\cdot)||_{\gamma-2,p}. \tag{2.7}$$

To estimate the right-hand side of (2.7), assume that $du = M^{-1}f\,dt + g^k\,dw^k_t$. Then

$$du_n(t,x) = f_n(t,x)\,dt + g_n(t,x)\,dw^k_t,$$

where $f_n(t,x) = (M^{-1}\zeta)(x)e^{-n}f(t,e^nx)$, $g_n(t,x) = \zeta(x)g(t,e^nx)$. By Lemma 2.12 with $c = e^{-np}$,

$$E\sup_{t\leq\tau}||u_{nx}(t,\cdot)||^p_{\gamma-2,p} \leq NT^{(p-2)/2}(e^{-np}||u_{nxx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau)}$$

$$+e^{np}||f_n||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} + ||g_{nx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau,l_2)}). \tag{2.8}$$

Furthermore, $||g_{nx}||_{H^{\gamma-2}_p(l_2)} \leq ||g_n||_{H^{\gamma-1}_p(l_2)}$ and

$$\sum_{n=-\infty}^{\infty} e^{n\theta}||g_n||^p_{\mathbb{H}^{\gamma-1}_p(\tau,l_2)} = ||g||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau,l_2)} \leq N||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}.$$

Also,

$$\sum_{n=-\infty}^{\infty} e^{n(\theta+p)}||f_n||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} = \sum_{n=-\infty}^{\infty} e^{n\theta}||f(\cdot,e^n\,\cdot)M^{-1}\zeta||^p_{\mathbb{H}^{\gamma-2}_p(\tau)}$$

$$\leq N||f||^p_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)} \leq N||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)},$$

$$\sum_{n=-\infty}^{\infty} e^{n(\theta-p)}||u_{nxx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} \leq \sum_{n=-\infty}^{\infty} e^{n(\theta-p)}||u_n||^p_{\mathbb{H}^{\gamma}_p(\tau)}$$

$$= \sum_{n=-\infty}^{\infty} e^{n(\theta-p)} ||(M^{-1}u)(\cdot, e^n \cdot)M\zeta||^p_{\mathbb{H}^\gamma_p(\tau)}$$

$$\leq N||M^{-1}u||^p_{\mathbb{H}^\gamma_{p,\theta}(\tau)} \leq N||u||^p_{\mathfrak{H}^\gamma_{p,\theta}(\tau)}.$$

By combining this with (2.8) and (2.6) we get (2.5). The theorem is proved.

As always the main role is played by the spaces $\mathfrak{H}^\gamma_{p,\theta,0}(\tau)$ of functions with zero as an initial condition. In connection with this it is worth noting that while constructing our theory we could replace

$$||u(0,\cdot)||^p_{U^\gamma_{p,\theta}} := E||M^{2/p-1}u(0,\cdot)||^p_{H^{\gamma-2/p}_{p,\theta}} \tag{2.9}$$

with

$$\inf\{||v_x||_{\mathbb{H}^{\gamma-1}_{p,\theta}} + ||\tilde{\mathbb{D}}v||_{\mathbb{H}^{\gamma-2}_{p,\theta}} + ||\tilde{\mathbb{S}}v||_{\mathbb{H}^{\gamma-1}_{p,\theta}} : u - v \in \mathfrak{H}^\gamma_{p,\theta,0}\}.$$

Such an axiomatic approach to defining a norm of $u(0,\cdot)$ yields, of course, the solvability results for the widest possible class of initial data, namely, for those which are extendible at least in some way for $t > 0$. However, in applications we often want to know how to describe "admissible" initial data by knowing only their analytic properties.

A partial answer to this question is given in the following theorem, which also shows why we use the norm given by (2.9). For the only case, which we need, $\gamma = 2$, the proof of this theorem can be obtained in the same way as Theorem 2.13 of [6]. For arbitrary γ and parabolic operators with coefficients depending only on time instead of Δ this theorem is proved in [5].

THEOREM 2.13. *If $\gamma \in \mathbb{R}$, $d-1 < \theta < p+d-1$, and $1 < p < \infty$, then, for every u_0 satisfying $u_0 \in U^\gamma_{p,\theta}$, in the space $\mathfrak{H}^\gamma_{p,\theta}$ there exists a unique solution of the heat equation $du = \Delta u\, dt$ with initial data $u(0,\cdot) = u_0$. Moreover,*

$$||u||_{\mathfrak{H}^\gamma_{p,\theta}} \leq N(d,\gamma,p,\gamma)||u_0||_{U^\gamma_{p,\theta}}.$$

3. SPDEs with constant coefficients in $\mathbb{R}^d_+$. Take a stopping time τ. On $[(0,\tau]] \cap (0,\infty)] \times \mathbb{R}^d_+$ we will be dealing with the following equation:

$$du = (a^{ij}u_{x^ix^j} + M^{-1}f)\, dt + (\sigma^{ik}u_{x^i} + g^k)\, dw^k_t \tag{3.1}$$

with initial condition $u|_{t=0} = u_0$, where u_0 is a $\mathcal{D}(\mathbb{R}^d_+)$-valued, $\mathcal{F}_0$-measurable random variable, f and g^k are $\mathcal{D}(\mathbb{R}^d_+)$-valued $\mathcal{P}$-measurable functions, a^{ij} and σ^{ik} are real-valued $\mathcal{P}$-measurable functions, u is an unknown $\mathcal{D}(\mathbb{R}^d_+)$-valued function, and the equation is understood in the sense of distributions as follows. We say that u is a solution of (3.1) with initial data u_0 if for any test function $\phi \in C^\infty_0(\mathbb{R}^d_+)$ we have

$$\begin{aligned}(u(t\wedge\tau,\cdot),\phi) &= (u_0,\phi)\\ &+ \int_0^{t\wedge\tau} \big[\sum_{i,j=1}^d a^{ij}(s)(u(s,\cdot),\phi_{x^ix^j}) + (f(s,\cdot),M^{-1}\phi)\big]\, ds\\ &+ \sum_{k=1}^\infty \int_0^{t\wedge\tau} \big[-\sum_{i=1}^d \sigma^{ik}(s)(u(s,\cdot),\phi_{x^i}) + (g^k(s,\cdot),\phi)\big]\, dw^k_s \qquad (3.2)\end{aligned}$$

for all $t > 0$ with probability one, where all integrals are assumed to have sense and the last series is assumed to converge uniformly on each interval of time $[0, T]$ in probability, where T is any finite constant.

Remark 3.1. If a function u belongs to $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, then it satisfies (3.1) with

$$f = M\left(\tilde{\mathbb{D}}u - a^{ij}D_iD_ju\right), \tag{3.3}$$

$$g^k = \tilde{\mathbb{S}}^k u - \sigma^{ik}D_iu.$$

In addition (see Lemma 1.3), we have $f \in \mathbb{H}^{\gamma-2}_{p,\theta}(\tau)$ and $g \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau, l_2)$. Below we show that under additional assumptions on θ, a, and σ the mapping $u \to (f, g)$ is onto.

Assumption 3.2. There exist constants $\delta_0, \delta_1 \in (0, 1]$ such that, for every (ω, t) and every $\xi \in \mathbb{R}^d$,

$$\delta_0|\xi|^2 \le \delta_1 a^{ij}(t)\xi^i\xi^j \le \bar{a}^{ij}\xi^i\xi^j \le a^{ij}(t)\xi^i\xi^j \le \delta_0^{-1}|\xi|^2,$$

where

$$\bar{a}^{ij} := a^{ij}(t) - \alpha^{ij}(t), \quad \alpha^{ij}(t) = \tfrac{1}{2}\sigma^{ik}(t)\sigma^{jk}(t).$$

Here is our main result.

THEOREM 3.3. *Let $d - 1 < \theta < p + d - 1$, $2 \le p < \infty$, $\gamma \in \mathbb{R}$, $f \in \mathbb{H}^{\gamma-2}_{p,\theta}(\tau)$, $g \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau, l_2)$, and $u_0 \in U^{\gamma}_{p,\theta}$. Assume that*

$$d - 1 + p\big[1 - \frac{1}{p(1-\delta_1) + \delta_1}\big] < \theta < d - 1 + p. \tag{3.4}$$

Then (3.1) *or equivalently* (3.3) *with initial data u_0 has a unique solution in $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$. In addition, for this solution it holds that*

$$||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)} \le N\big(||f||^p_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)} + ||g||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau,l_2)} + ||u_0||^p_{U^{\gamma}_{p,\theta}}\big), \tag{3.5}$$

where $N = N(\gamma, \theta, p, d, \delta_0, \delta_1)$.

Remark 3.4. By Remark 2.7, one gets a statement equivalent to Theorem 3.3 if one replaces $M^{-1}f$ in (3.1) with D_if^i for certain $f^i \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$ and replaces $||f||^p_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)}$ in (3.5) with $\sum_i ||f^i||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)}$.

Remark 3.5. If $\sigma \equiv 0$, then one can take $\delta_1 = 1$ and (3.4) becomes $d - 1 < \theta < d - 1 + p$. Furthermore, it is easy to see that, for any σ, condition (3.4) is satisfied if $d - 2 + p \le \theta < d - 1 + p$.

Remark 3.6. It is worth noting that if $\theta \ge p + d - 1$ or $\theta \le d - 1$, then the statement of Theorem 3.3 is false even in the case of the heat equation. This can be shown by simple examples.

The proof of this theorem is based on two lemmas, the first of which we prove in section 4.

LEMMA 3.7. *Theorem* 3.3 *holds if $\gamma = 2$.*

LEMMA 3.8. *Let the assumptions of Theorem* 3.3 *be satisfied and let $\mu \le \gamma$. Let $\theta_1 \in \mathbb{R}$ and let $u \in \mathfrak{H}^{\mu}_{p,\theta_1}(\tau)$ be a solution of* (3.1) *with initial condition u_0. Assume that $M^{-1}u \in \mathbb{H}^{\mu}_{p,\theta}(\tau)$. Then $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ and*

$$||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)} \le N\big(||f||^p_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)} + ||g||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau,l_2)} + ||u_x||^p_{\mathbb{H}^{\mu-1}_{p,\theta}(\tau)} + ||u_0||^p_{U^{\gamma}_{p,\theta}}\big),$$

where $N = N(d, \gamma, \mu, \theta, p)$.

One can prove this lemma by repeating almost word for word the proof of Lemma 3.5 of [6]. The only noticeable difference is that the equations in [6] are written in the form

$$du = (au_{xx} + f_x)\,dt + (\sigma^k u_x + g^k)\,dw^k_t,$$

where we have f_x instead of $M^{-1}f$. But by Remark 3.4 we also can rewrite (3.1) with $D_i f^i$ in place of $M^{-1}f$.

Proof of Theorem 3.3. As in the proof of Theorem 3.2 of [6] we may assume that $\tau = \infty$. In the case $\gamma \geq 2$ the proof is achieved on the basis of Lemma 3.8 by repeating the proof of Theorem 3.2 of [6]. In the case $\gamma < 2$ we need only some minor adjustments which we present for completeness.

Denote by $\mathcal{R}$ the operator which maps (f, g, u_0) with $f \in \mathbb{H}^{\gamma-2}_{p,\theta}$, $g \in \mathbb{H}^{\gamma-1}_{p,\theta}(l_2)$, and $u_0 \in U^{\gamma}_{p,\theta}$ into the solution $u \in \mathfrak{H}^{\gamma}_{p,\theta}$ of (3.1) with initial data u_0. Thus far we know that $\mathcal{R}$ is well defined in spaces $\mathbb{H}^{\gamma-2}_{p,\theta} \times \mathbb{H}^{\gamma-1}_{p,\theta}(l_2) \times U^{\gamma}_{p,\theta}$ for $\gamma \geq 2$. We want to show that one can also define $\mathcal{R}$ for $\gamma < 0$.

First, let $2 > \gamma \geq 1$. Observe that by Lemma 1.3, part (iii),

$$(\mathcal{L}^{-1}f, \mathcal{L}^{-1}g, M^{1-2/p}\mathcal{L}^{-1}M^{2/p-1}u_0) \in \mathbb{H}^{\gamma}_{p,\theta} \times \mathbb{H}^{\gamma+1}_{p,\theta}(l_2) \times U^{\gamma+2}_{p,\theta}.$$

Since $\gamma > 0$, by what we know in the case $\gamma \geq 2$, the function

$$v = \mathcal{R}(\mathcal{L}^{-1}f, \mathcal{L}^{-1}g, M^{1-2/p}\mathcal{L}^{-1}M^{2/p-1}u_0)$$

is well defined and belongs to $\mathfrak{H}^{\gamma+2}_{p,\theta}$.

Define

$$\tilde{u} = \mathcal{L}v.$$

By Remark 2.8, we have $\tilde{u} \in \mathfrak{H}^{\gamma}_{p,\theta}$. Furthermore, by definition v satisfies

$$dv = (a^{ij}v_{x^ix^j} + M^{-1}\mathcal{L}^{-1}f)\,dt + (\sigma^{ik}v_{x^i} + \mathcal{L}^{-1}g^k)\,dw^k_t.$$

We apply $\mathcal{L}$ to both parts of this equality, or in other words, we substitute $\mathcal{L}^*\phi$ in place of ϕ in (3.2), where $\mathcal{L}^*$ is the formal adjoint to $\mathcal{L}$. Then we get

$$d\tilde{u} = (a^{ij}\tilde{u}_{x^ix^j} + M^{-1}f + M^{-1}\bar{f})\,dt + (\sigma^{ik}\tilde{u}_{x^i} + g^k + \bar{g}^k)\,dw^k_t,$$

$$\tilde{u}(0, \cdot) = u_0 + \bar{u}_0,$$

where

$$\bar{f} = M\mathcal{L}a^{ij}v_{x^ix^j} - Ma^{ij}(\mathcal{L}v)_{x^ix^j} + M\mathcal{L}M^{-1}\mathcal{L}^{-1}f - f,$$

$$\bar{g}^k = \mathcal{L}\sigma^{ik}v_{x^i} - \sigma^{ik}(\mathcal{L}v)_{x^i}, \quad \bar{u}_0 = \mathcal{L}M^{1-2/p}\mathcal{L}^{-1}M^{2/p-1}u_0 - u_0.$$

Next, we use

$$\mathcal{L}D_i\phi = D_i\mathcal{L}\phi - 2\delta^{i1}M^{-1}(\mathcal{L} - MD_1)\phi,$$

$$\mathcal{L}M^{-1}\phi = M^{-1}\mathcal{L}\phi - 2D_1\phi.$$

Then we find that

$$\bar{f} = -2a^{i1}MD_iM^{-1}(\mathcal{L} - MD_1)v - 2a^{1j}(\mathcal{L} - MD_1)D_jv - 2MD_1\mathcal{L}^{-1}f,$$

$$\bar{g}^k = -2\sigma^{1k}M^{-1}(\mathcal{L} - MD_1)v,$$

$$M^{2/p-1}\bar{u}_0 = (2 - 4/p)MD_1\mathcal{L}^{-1}M^{2/p-1}u_0 + c\mathcal{L}^{-1}M^{2/p-1}u_0,$$

where c is a constant. As above

$$(\mathcal{L} - MD_1)v \in \mathfrak{H}^\gamma_{p,\theta}, \quad M^{-1}(\mathcal{L} - MD_1)v \in \mathbb{H}^\gamma_{p,\theta}, \quad Dv \in \mathbb{H}^{\gamma+1}_{p,\theta},$$

$$M^{2/p-1}\bar{u}_0 \in L_p(\Omega, \mathcal{F}_0, H^{\gamma+1-2/p}_{p,\theta}).$$

It follows that

$$(\bar{f}, \bar{g}, \bar{u}_0) \in \mathbb{H}^{\gamma-1}_{p,\theta} \times \mathbb{H}^\gamma_{p,\theta}(l_2) \times U^{\gamma+1}_{p,\theta}. \tag{3.6}$$

Since $\gamma \geq 1$, it follows from (3.6) that the function $\bar{u} := \mathcal{R}(\bar{f}, \bar{g}, \bar{u}_0)$ is well defined, belongs to $\mathfrak{H}^{\gamma+1}_{p,\theta}$, and the function $u = \tilde{u} - \bar{u}$ is of class $\mathfrak{H}^\gamma_{p,\theta}$ and solves (3.1). For thus constructed u, estimate (3.5) follows from the explicit representation and known estimates for $\mathcal{R}$, $\mathcal{L}$, MD.

By repeating the above argument, we consider the case $1 > \gamma \geq 0$, this time using the fact that $\gamma + 1 \geq 1$ and relying upon the result for $\gamma \geq 1$. One can continue in the same way and it remains to prove only the uniqueness of solutions in $\mathfrak{H}^\gamma_{p,\theta}$.

It suffices to consider the case $f = 0, g = 0, u_0 = 0$ (and $\gamma < 2$). In this case any solution $u \in \mathfrak{H}^\gamma_{p,\theta,0}$ also belongs to $\mathfrak{H}^2_{p,\theta,0}$ by Lemma 3.8 and its uniqueness follows from Lemma 3.7.

The theorem is thus proved.

Remark 3.9. From the above derivation of Theorem 3.3 from Lemma 3.7 it is seen that for *any* fixed $\gamma, p, \theta, a, \sigma$ satisfying the conditions of Theorem 3.3, if the assertion of Theorem 3.3 holds for these $\gamma, p, \theta, a, \sigma$, then it holds for any $\gamma \in \mathbb{R}$ with the same p, θ, a, σ.

4. Proof of Lemma 3.7. By Remarks 2.7, we may concentrate on the following form of (3.1):

$$\begin{aligned} du(t,x) = (a^{ij}(t)u_{x^ix^j}(t,x) + D_if^i(t,x))dt \\ + (\sigma^{ik}(t)u_{x^i}(t,x) + g^k(t,x))dw^k(t). \end{aligned} \tag{4.1}$$

Next, notice that by Theorem 2.13 there is a function $\bar{u} \in \mathfrak{H}^2_{p,\theta}$ such that, $\bar{u}|_{t=0} = u_0$, $\partial\bar{u}/\partial t = D_i\bar{f}^i$ with $\bar{f} \in \mathbb{H}^1_{p,\theta}$, and appropriate estimates of $||\bar{u}_x||_{\mathbb{H}^1_{p,\theta}}$ and $||\bar{f}||_{\mathbb{H}^1_{p,\theta}}$ through $||u_0||_{U^2_{p,\theta}}$ hold. This implies that in the equation

$$du = (a^{ij}u_{x^ix^j} + (a^{ij}\bar{u}_{x^j} + f^i - \bar{f}^i)_{x^i})\,dt + (\sigma^{ik}u_{x^i} + (\sigma^{ik}\bar{u}_{x^i} + g^k))\,dw^k_t$$

we have $a^{ij}\bar{u}_{x^j} + f^i - \bar{f}^i \in \mathbb{H}^1_{p,\theta}(\tau)$ and $\sigma^{i\cdot}\bar{u}_{x^i} + g \in \mathbb{H}^1_{p,\theta}(\tau, l_2)$. Also, obviously if we can solve the above equation in $\mathfrak{H}^2_{p,\theta,0}(\tau)$, then by adding to the solution the function

$\bar{u}$ we get a solution of (4.1) with initial data u_0. Therefore, in the proof of Lemma 3.7 without loss of generality, we may and will confine ourselves only to the case $u_0 \equiv 0$.

Finally, obviously we may assume that $\tau \leq T$, where the constant $T < \infty$, and we start by proving the following a priori estimate.

LEMMA 4.1. *Assume that there exists a constant $\delta_2 > 0$ such that*

$$(p-1)(d+p-1-\theta)\bar{a}^{11} - p(d+p-2-\theta)a^{11} \geq \delta_2 \tag{4.2}$$

for all ω and t. Then for any $u \in \mathfrak{H}^2_{p,\theta,0}(\tau)$,

$$||M^{-1}u||_{\mathbb{L}_{p,\theta}(\tau)} \leq N(||M(\tilde{\mathbb{D}} - a^{ij}D_iD_j)u||_{\mathbb{L}_{p,\theta}(\tau)} + ||(\tilde{\mathbb{S}} - \sigma^{i\cdot}D_i)u||_{\mathbb{L}_{p,\theta}(\tau,l_2)}), \tag{4.3}$$

where N depends only on $\delta_0, \delta_2, d,$ and p.

Proof. For any γ the operators $M\tilde{\mathbb{D}}$ and $\tilde{\mathbb{S}}$ are obviously continuous on $\mathfrak{H}^\gamma_{p,\theta}(\tau)$ with values in $\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)$ and $\mathbb{H}^{\gamma-1}_{p,\theta}(\tau, l_2)$, respectively. By Remark 2.5 the same is true for $M^{-1} : \mathfrak{H}^\gamma_{p,\theta}(\tau) \to \mathbb{H}^\gamma_{p,\theta}(\tau)$. By Definition 2.2 and Lemma 1.3 the operators

$$MD_iD_j : \mathfrak{H}^\gamma_{p,\theta}(\tau) \to \mathbb{H}^{\gamma-2}_{p,\theta}(\tau), \quad \sigma^{ik}D_i : \mathfrak{H}^\gamma_{p,\theta}(\tau) \to \mathbb{H}^{\gamma-1}_{p,\theta}(\tau, l_2)$$

are bounded. By Theorem 2.9, it follows that we need to prove only (4.3) for functions u belonging to set (2.3) with sufficiently large n.

Take such a function u and define f and g according to (3.3). By Sobolev's embedding theorem, if n is large, then f and g are continuous in x, have compact supports in x, and

$$E\int_0^\tau \sup_x |f(t,x)|^p\,dt < \infty, \quad E\int_0^\tau \sup_x |g(t,x)|^p_{l_2}\,dt < \infty.$$

It follows easily that u satisfies (3.1) pointwise, that is, for almost any ω for all $x \in \mathbb{R}^d_+$ and $t \in [0, \tau]$.

Next we define $c = 2 + \theta - d - p$, apply Itô's formula to $(x^1)^c|u(t,x)|^p$, and find almost surely for all $x \in \mathbb{R}^d_+$

$$(x^1)^c|u(\tau,x)|^p = \int_0^\tau \Bigg[p(x^1)^c|u|^{p-2}ua^{ij}u_{x^ix^j}$$

$$+p(x^1)^{c-1}|u|^{p-2}uf + \tfrac{1}{2}p(p-1)(x^1)^c|u|^{p-2}\sum_k(\sigma^{ik}u_{x^i} + g^k)^2\Bigg]\,ds$$

$$+\int_0^\tau p(x^1)^c|u|^{p-2}u(\sigma^{ik}u_{x^i} + g^k)\,dw^k_s. \tag{4.4}$$

We take expectations of both parts of this equality, noticing that

$$E\left[\int_0^\tau |u|^{2p-2}\sum_k |\sigma^{ik}u_{x^i} + g^k|^2\,ds\right]^{1/2} \tag{4.5}$$

$$\leq NTE\sup_{s\leq\tau}|u|^{p-1}|u_x| + NE\sup_{s\leq\tau}|u|^{p-1}\left[\int_0^\tau |g|^2_{l_2}\right]^{1/2}.$$

Here, for instance, by Hölder's inequality the last expectation is less than

$$\big(E\sup_{s\leq\tau}|u|^p\big)^{(p-1)/p}\left(T^{(p-2)/2}E\int_0^\tau |g|_{l_2}^p\,ds\right)^{1/p}<\infty.$$

Therefore, the left-hand side of (4.5) is finite and the stochastic integral will disappear after taking expectations in (4.4). After this we integrate with respect to x over $\mathbb{R}^d_+$. By the way, owing to the fact that x-supports of all functions u, f, and g belong to some G_k and the fact that even the pth power of sup's over x of these functions are integrable over $(0,\tau]\!]$, we see that all integrals converge absolutely. Hence, by using Fubini's theorem and integrating by parts, we get from (4.4) that

$$0\leq E\int_0^\tau\int_{\mathbb{R}^d_+}\big[-p(p-1)(x^1)^c|u|^{p-2}\bar{a}^{ij}u_{x^i}u_{x^j}$$

$$-c(x^1)^{c-1}a^{i1}(|u|^p)_{x^i}+p(p-1)(x^1)^c|u|^{p-2}g^k\sigma^{ik}u_{x^i}$$

$$+p(x^1)^{c-1}|u|^{p-1}|f|+\tfrac{1}{2}p(p-1)(x^1)^c|u|^{p-2}|g|_{l_2}^2\big]\,dxdt.$$

Next, we use Young's inequality to get relations like

$$(x^1)^{c-1}|u|^{p-1}|f|\leq\varepsilon(x^1)^{\theta-d}|u/x^1|^p+N(x^1)^{\theta-d}|f|^p,$$

$$g^k\sigma^{ik}u_{x^i}\leq N|g|_{l_2}|u_x|\leq\varepsilon\bar{a}^{ij}u_{x^i}u_{x^j}+N|g|_{l_2}^2,$$

where $\varepsilon>0$ is arbitrary and N depends only on ε, δ_0, and p. Then we get

$$0\leq E\int_0^\tau\int_{\mathbb{R}^d_+}\big[(\varepsilon-p(p-1))(x^1)^c|u|^{p-2}\bar{a}^{ij}u_{x^i}u_{x^j}$$

$$+(\varepsilon+c(c-1))a^{11}(x^1)^{\theta-d}|u/x^1|^p+N(x^1)^{\theta-d}|f|^p+N(x^1)^{\theta-d}|g|_{l_2}^p\big]\,dxdt.$$

By Corollary 6.2 of [5] for any t

$$\int_{\mathbb{R}^d_+}(x^1)^c|u|^{p-2}\bar{a}^{ij}u_{x^i}u_{x^j}\geq\bar{a}^{11}(1-c)^2p^{-2}\int_{\mathbb{R}^d_+}(x^1)^{\theta-d}|u/x^1|^p\,dx.$$

Hence,

$$E\int_0^\tau\{\bar{a}^{11}[p(p-1)-\varepsilon](1-c)^2p^{-2}+a^{11}[c(1-c)-\varepsilon]\}||M^{-1}u||_{0,p,\theta}^p\,dt$$

$$\leq N(||f||^p_{\mathbb{L}_{p,\theta}(\tau)}+||g||^p_{\mathbb{L}_{p,\theta}(\tau,l_2)}).$$

It remains only to observe that for ε small enough from (4.2) we get that

$$\bar{a}^{11}[p(p-1)-\varepsilon](1-c)^2p^{-2}+a^{11}[c(1-c)-\varepsilon]$$

$$\geq -(1-c)p^{-1}\delta_2/2 + \bar{a}^{11}(p-1)(1-c)^2p^{-1} + a^{11}c(1-c)$$

$$= -(1-c)p^{-1}\delta_2/2 + (1-c)p^{-1}[(p-1)(d+p-1-\theta)\bar{a}^{11} - p(d+p-2-\theta)a^{11}]$$

$$\geq (1-c)p^{-1}\delta_2/2.$$

The lemma is proved.

We divide the remaining part of the proof of Lemma 3.7 into the following subcases:

(1) $\sigma \equiv 0$;

(2) general case.

4.1. Case $\boldsymbol{\sigma \equiv 0}$. Observe that in this case $\bar{a} = a$ and (4.2) becomes

$$a^{11}(\theta - d + 1) \geq \delta_2,$$

which is satisfied for δ_2 sufficiently small because we always assume that $\theta > d-1$ (and, for that matter, $\theta < p+d-1$). Therefore, estimate (4.3) holds. Of course, this estimate implies uniqueness.

To prove existence again use (4.3) and proceed as in the proof of Lemma 4.2 of [6] or Lemma 5.7 of [5]. Since this can be done in quite a straightforward way, we give only a sketch.

First, bearing in mind the a priori estimate and the method of continuity, we see that it suffices to consider the case $a^{ij} = \delta^{ij}$. Furthermore, owing to Theorem 2.10 and Lemma 4.1 we may and will additionally assume that

$$f \in L_p(\Omega, C((0,\tau], C_0^n(G_k))), \quad g \in L_p(\Omega, C((0,\tau], C_0^n(G_k))).$$

Continue f and g across $x^1 = 0$ so that f becomes an even smooth function and g an odd smooth function of x^1. By Theorem 3.2 of [3] or Theorem 5.1 of [4] there exists a unique solution $u \in \mathcal{H}_p^n(\tau)$ of (3.1) considered in the whole $\mathbb{R}^d$ with zero initial condition. If n is large enough, u is smooth with respect to x and satisfies (3.1) pointwise. From the uniqueness, it follows that $u(t,x) = 0$ for $x^1 = 0$. Next use the fact that the functions f and g have compact support and that outside this support u satisfies the deterministic equation $du = \Delta u\,dt$. Then as in the proof of Lemma 4.2 of [6] we derive that $u \in \mathfrak{H}^2_{p,\theta,0}(\tau)$. Using Lemma 3.8 with $\gamma = 2$ and $\mu = 0$ and Lemma 4.1 we conclude that u belongs to $\mathfrak{H}^2_{p,\theta,0}(\tau)$, satisfies (3.1), and estimate (3.5) holds for $\gamma = 2$ and $u_0 = 0$. This proves Lemma 3.7 in our first particular case.

4.2. General case. The left inequality in (3.4) means that

$$\delta_1(p-1)(d+p-1-\theta) > p(d+p-2-\theta),$$

which by virtue of Assumption 3.2 implies (4.2) with

$$\delta_2 = \delta_0[\delta_1(p-1)(d+p-1-\theta) - p(d+p-2-\theta)] > 0.$$

Therefore, a priori estimate (4.3) holds. Using Lemma 3.8 with $\gamma = 2$ and $\mu = 0$, we get that estimate (3.5) holds for $\gamma = 2$ and $u_0 = 0$. In particular, we get the uniqueness.

Furthermore, the same estimate with the same constant N holds if we take $\lambda\sigma^{ik}$ instead of σ^{ik} if $|\lambda| \leq 1$. Now to get the result in our present case from the case $\sigma \equiv 0$ it remains only to use the method of continuity (cf., for instance, the end of the proof of Theorem 5.1 of [4]).

REFERENCES

[1] N. V. Krylov, *A W_2^n-theory of the Dirichlet problem for SPDEs in general smooth domains*, Probab. Theory Related Fields, 98 (1994), pp. 389–421.

[2] N. V. Krylov, *A generalization of the Littlewood–Paley inequality and some other results related to stochastic partial differential equations*, Ulam Quart., 2 (1994), pp. 16–26; also available online from http://www.ulam.usm.edu/VIEW2.4/krylov.ps.

[3] N. V. Krylov, *On L_p-theory of stochastic partial differential equations in the whole space*, SIAM J. Math. Anal., 27 (1996), pp. 313–340.

[4] N. V. Krylov, *An analytic approach to SPDEs*, in Stochastic Partial Differential Equations, Six Perspectives, Math. Surveys Monog., AMS, Providence, RI, 1999, pp. 185–242.

[5] N. V. Krylov, *Weighted Sobolev spaces and Laplace's and the heat equations in a half space*, Comm. Partial Differential Equations, 24 (1999), pp. 1611–1653.

[6] N. V. Krylov and S. V. Lototsky, *A Sobolev space theory of SPDEs with constant coefficients on a half line*, SIAM J. Math. Anal., 30 (1999), pp. 298–325.

[7] S. Lapic, *On the First Initial-Boundary Problem for SPDEs on Domains with Limited Smoothness at the Boundary*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1994.

[8] B. L. Rozovskii, *Stochastic Evolution Systems*, Kluwer, Dordrecht, The Netherlands, 1990.

[9] H. Triebel, *Theory of Function Spaces* II, Birkhäuser Verlag, Basel, 1992.

SIAM J. Math. Anal.
Vol. 31, No. 1, pp. 34–62

CONTINUOUS DEPENDENCE IN CONSERVATION LAWS WITH PHASE TRANSITIONS*

RINALDO M. COLOMBO[†] AND ANDREA CORLI[‡]

Abstract. This paper is concerned with systems of 2×2 conservation laws

$$(\star) \qquad \partial_t u + \partial_x [f(u)] = 0, \qquad t \geq 0\,,\ x \in \mathbf{R}\,,\ u \in \mathbf{R}^2,$$

developing phase transitions, as happens in models related to elastodynamics or to van der Waals fluids, for instance.

In the present paper, a definition of Ψ-admissible solution to $(\star)$ is given which comprises the various definitions in the current literature. Furthermore, the Ψ-admissible Riemann semigroup (ΨRS) generated by $(\star)$ is introduced and constructed by means of a wave-front tracking algorithm. Uniqueness and continuous dependence for Ψ-admissible solutions to $(\star)$ thus follow.

Key words. conservation laws, phase boundaries

AMS subject classifications. 35L65, 35L67

PII. S0036141097331871

1. Introduction.

This paper is concerned with phase transitions in systems of 2×2 conservation laws

$$\partial_t u + \partial_x [f(u)] = 0, \tag{1.1}$$

where $t \geq 0$, $x \in \mathbf{R}$, $u \in \Omega$, and $\Omega \subset \mathbf{R}^2$. The smooth flow function $f\colon \Omega \mapsto \mathbf{R}^2$ is strictly hyperbolic. We assume that Ω is the union of two disjoint open sets Ω_1 and Ω_2 which we refer to as *phases*. The eigenvalues of Df are assumed to be either genuinely nonlinear or linearly degenerate in each phase. By *phase transition* we mean a jump in a solution to (1.1) whose states on the two sides of the discontinuity belong to different phases.

Phase transitions model abrupt changes in some physical property of the system under consideration. A well-known example comes from elastodynamics, where $u = (v, w)$ and $f(u) = (-\sigma(w), -v)$. Here v is the particle speed, w is the strain, and σ is a nonmonotone stress-strain function. In recent years many authors dealt with this model; we quote for brevity only [11], [1], [14], [13]. Another model developing phase boundaries is the system of van der Waals fluids, as considered in [18], [19], [8]. We refer the reader to the survey paper [21] and to [23] for other physical accounts.

The introduction of phase transitions in (1.1) may be necessary in order to solve the Riemann problem

$$\begin{cases} \partial_t u + \partial_x [f(u)] = 0, \\ u(0, x) = u_o(x), \end{cases} \qquad \text{where} \qquad u_o(x) = \begin{cases} u^\flat & \text{if } x < 0, \\ u^\sharp & \text{if } x > 0 \end{cases} \tag{1.2}$$

*Received by the editors December 23, 1997; accepted for publication (in revised form) October 6, 1998; published electronically November 4, 1999. This work was partially supported by TMR project HCL ERBFMRXCT960033 of the European Union.

http://www.siam.org/journals/sima/31-1/33187.html

[†] University of Milano, Department of Mathematics, Via Saldini 50, 20133 Milano, Italy (rinaldo@ing.unibs.it). Current address: Department of Mathematics, University of Brescia, Via Branze 38, 25123 Brescia, Italy.

[‡] University of Ferrara, Department of Mathematics, Via Machiavelli 35, 44100 Ferrara, Italy (crl@dns.unife.it).

with $u^\flat, u^\sharp$ in the same phase but not necessarily close. It may well happen that no physically acceptable solution to (1.2) exists, unless a middle state $u^\natural$ is chosen in the other phase. From a mathematical point of view, the Lax shock-rarefaction curves may have no intersection in the phase which contains $u^\flat$ and $u^\sharp$. Nonetheless, a physically acceptable solution can be defined by introducing two phase boundaries between $u^\flat$ and $u^\sharp$.

We shall consider only *subsonic* phase transitions, which means that the absolute value of the speed of the discontinuity is lower than the absolute value of the characteristic speeds. An important feature of this case is that the Rankine–Hugoniot conditions turn out to be insufficient to uniquely determine a solution to (1.2) and a further *admissibility* condition, expressed by a function Ψ, is required.

From our point of view, we are not interested in the particular admissibility condition that is added, provided it satisfies some minimal regularity and stability assumptions. We consider it to be a *physical* problem to select the most suitable admissibility condition for every single specific application of (1.1). We remark, however, that our procedures apply to both the kinetic approach of elastodynamics [1] and the visco-capillarity approach [18], [19], [8] of van der Waals' model.

The main result of this paper is the construction of a Ψ-admissible Riemann semigroup (ΨRS) whose orbits are solutions to

$$\begin{cases} \partial_t u + \partial_x \left[f(u) \right] = 0, \\ u(0,x) = \underline{u}(x) + \tilde{u}(x), \end{cases} \tag{1.3}$$

where $\tilde{u}$ is assumed to have suitably small total variation. $\underline{u}$ is the solution to (1.2) evaluated at some nonnegative time, with $u^\flat$, $u^\sharp$ in the same or in different phases. For example, the case $\underline{u} = u_o$ is acceptable.

The main tool is a modification of the wave-front tracking algorithm as developed in [4], [5] for the construction of a standard Riemann semigroup (SRS). We refer to [3] for a review of the SRS theory.

The above results depend on a *stability* and a *strong nonresonance* condition on the unperturbed problem (1.2). The former was first stated in [6]; the latter resembles what is done in [5]. A notable difference with respect to [5] is the addition of a condition to single out the solutions to the Riemann problems, as we mentioned above; moreover, the generic interaction of a wave against a phase boundary leads to a configuration entirely different from the case considered therein.

Due to the result in this paper, the whole recent theory of SRSs [3] and *viscosity solutions* [2] can be extended to systems of the form (1.1) that develop phase boundaries. We recall only that by means of the definition of viscosity solutions the trajectories of the SRS are characterized in terms of integral inequalities relying solely on (1.1). Furthermore, by introducing a condition analogous to (A3) in [5], it is possible to uniquely characterize the solution constructed here, thus providing an *existence and uniqueness* theorem for viscosity solution to (1.1) satisfying (A3) and developing admissible phase boundary.

The paper is organized as follows. In section 2 we collect some basic facts, give precise definitions, and state our main result. The main theorem is applied in section 3 to the problem of elastodynamics and to a model of a van der Waals gas. Section 4 contains the statements of a number of propositions, which are proved in the last section 5.

2. Notations and main results. Let Ω be the union of two disjoint open subsets Ω_1, Ω_2 of $\mathbf{R}^2$; Ω_1 and Ω_2 are called *phases*. Throughout this paper we assume that $f\colon \Omega \mapsto \mathbf{R}^2$ is a smooth function and that its Jacobian matrix $A(u) = Df(u)$ is strictly hyperbolic in all Ω; this means that $A(u)$ has two real and distinct eigenvalues $\lambda_1(u)$, $\lambda_2(u)$ for every $u \in \Omega$. By eventually applying a linear change of coordinates, we assume that

$$(2.1) \qquad -\lambda^{\max} < \lambda_1(u) < -\lambda^{\min} < 0 < \lambda^{\min} < \lambda_2(u) < \lambda^{\max} \qquad \text{for all } u \in \Omega$$

for two fixed constants $\lambda^{\min}$, $\lambda^{\max}$. We denote by $r_1(u)$, $r_2(u)$ the eigenvectors associated to the eigenvalues $\lambda_1(u)$, $\lambda_2(u)$. Each characteristic family is locally either linearly degenerate or genuinely nonlinear, i.e., we assume that

$$\text{either} \quad \nabla\lambda_i(u) \cdot r_i(u) = 0 \quad \text{or} \quad \nabla\lambda_i(u) \cdot r_i(u) > 0 \quad \text{for all } u \in \mathcal{U}^\flat$$

and similarly for $\mathcal{U}^\natural$ and $\mathcal{U}^\sharp$. Here $\mathcal{U}^\flat$, $\mathcal{U}^\natural$, and $\mathcal{U}^\sharp$ are subsets of Ω which are going to be specified in the following.

The fact that Ω is the disjoint union of two open sets has important consequences on the solutions of the Riemann problem (1.2). First, if $u^\flat$ and $u^\sharp$ belong to different phases, no Lax [12] solution can be defined. But even if $u^\flat$ and $u^\sharp$ both belong to the same phase Ω_j, then again the Lax solution may not exist. In fact, the shock-rarefaction curves through $u^\flat$, $u^\sharp$ may have no intersection inside Ω_j. In these cases, *phase boundaries* arise.

More precisely, let $u\colon [0, +\infty[\times \mathbf{R} \mapsto \Omega$ be a weak solution to (1.1) such that $u(t, \cdot) \in BV$ for all t. A Lipschitz continuous curve $x = \Lambda(t)$ is a *phase boundary* for u if for almost every t the traces

$$u^l(t) = \lim_{x \to \Lambda(t)-} u(t,x) \qquad \text{and} \qquad u^r(t) = \lim_{x \to \Lambda(t)+} u(t,x)$$

are in two different phases. When this happens, the Rankine–Hugoniot condition

$$(2.2) \qquad \dot\Lambda \cdot \left(u^l - u^r\right) = f(u^l) - f(u^r)$$

must be satisfied for a.e. t in order to have a weak solution. By eliminating $\dot\Lambda$ in (2.2), the Rankine–Hugoniot equations reduce to the scalar condition

$$(2.3) \qquad \Phi_{RH}(u^l, u^r) = 0$$

for a suitable smooth function Φ_{RH}.

In the rest of this paper, we consider only *subsonic* phase boundaries, i.e., we assume that

$$(2.4) \qquad \left|\dot\Lambda\right| < \lambda^{\min} .$$

The choice (2.4) is motivated by the fact that phase boundaries satisfying $\left|\dot\Lambda\right| > \lambda^{\max}$ can be treated as overcompressive shocks [16], but this situation does not seem physically relevant; see [11]. In the intermediate *supersonic* case $\lambda_i(u^l) < \dot\Lambda < \lambda_i(u^r)$ (or $\lambda_i(u^r) < \dot\Lambda < \lambda_i(u^l)$), Lax shock inequalities are satisfied and phase boundaries behave as large shocks and can be treated exactly as in [5]. For sonic phase boundaries, we refer to [7].

It is possible to define a solution to the Riemann problem (1.2) relying solely on (2.3). However, in the subsonic case (2.4) the requirement (2.3) alone does not single out a unique solution. It is thus necessary to impose a further constraint on the states on the sides of the discontinuity, say,

$$\Psi(u^l, u^r) = 0, \tag{2.5}$$

where $\Psi\colon (\Omega_1 \times \Omega_2) \cup (\Omega_2 \times \Omega_1) \to \mathbf{R}$ is a smooth function. The conditions (2.3) and (2.5) allow us to single out a unique solution to (1.2). We shall consider only Ψ-*admissible* phase boundaries, i.e., those whose side states satisfy (2.3) and (2.5). We stress that our condition depends on the particular choice of Ψ.

From a physical point of view, the choice of the function Ψ is usually related to some kind of entropy dissipation, as in the examples developed in section 3. Let us point out that the case of an admissibility function Ψ depending solely on the speed of the phase boundary is contained in our framework, since the jump conditions (2.3) allow us to express $\dot{\Lambda}$ by means of u^l and u^r.

Under assumption (2.4) and having imposed condition (2.5), it is natural to give the following definition.

Definition 2.1. *We call Ψ-admissible solution to the Riemann problem* (1.2) *under the admissibility condition* (2.5)

(i) *the usual Lax solution as long as it exists;*

(ii) *the solution consisting of a Lax wave of the first family, a Ψ-admissible phase boundary, and a Lax wave of the second family, whenever $u^\flat$ and $u^\sharp$ belong to different phases;*

(iii) *the solution consisting of a Lax wave of the first family, two Ψ-admissible phase boundaries, and a Lax wave of the second family, whenever $u^\flat$ and $u^\sharp$ belong to the same phase but a Lax solution does not exist.*

For notational simplicity, we introduce the function $\Phi\colon (\Omega_1 \times \Omega_2) \cup (\Omega_2 \times \Omega_1) \to \mathbf{R}^2$ by

$$\Phi(u^l, u^r) = \begin{bmatrix} \Phi_{RH}(u^l, u^r) \\ \Psi(u^l, u^r) \end{bmatrix} .$$

Denote by $D_1\Phi$ (resp., $D_2\Phi$) the 2×2 Jacobian of Φ with respect to the first (resp., second) argument. It is useful to consider the following situations separately:

(1) $\Phi(u^\flat, u^\sharp) = 0$, so that (1.2) is solved by a subsonic phase boundary and no other waves;

(2) $\Phi(u^\flat, u^\natural) = 0$ and $\Phi(u^\natural, u^\sharp) = 0$ for some middle state $u^\natural$, so that (1.2) is solved by two subsonic phase boundaries and no other waves.

We emphasize that we do not take into account in this paper the phenomenon of *nucleation*, i.e., the initiation of two phase boundaries at a certain time from data in the same phase (see [1], [13]). To prevent nucleation we shall make some assumptions in order that the solutions to (1.3) have the same number of phase boundaries of (1.2), which is, according to the previous situations, either one or two.

In case (1), we say that the phase boundary separating $u^\flat$ from $u^\sharp$ is *stable* when

$$\det\Big(D_1\Phi(u^\flat, u^\sharp) r_1(u^\flat),\ D_2\Phi(u^\flat, u^\sharp) r_2(u^\sharp)\Big) \neq 0. \tag{2.6}$$

This condition ensures the unique solvability in the sense of Definition 2.1, part (ii), of all Riemann problems with data sufficiently near to $u^\flat$, $u^\sharp$ by the implicit function theorem. Similarly, in case (2), we assume that

$$\det\Big(D_1\Phi(u^\flat, u^\natural) r_1(u^\flat),\ D_2\Phi(u^\flat, u^\natural) r_2(u^\natural)\Big) \neq 0, \tag{2.7}$$

$$\det\Big(D_1\Phi(u^\natural, u^\sharp) r_1(u^\natural),\ D_2\Phi(u^\natural, u^\sharp) r_2(u^\sharp)\Big) \neq 0. \tag{2.8}$$

The conditions above imply the solvability of small perturbations of the Riemann problems with data $(u^\flat, u^\natural)$ and $(u^\natural, u^\sharp)$. Due to Definition 2.1, we need a further global condition ensuring that no small perturbation of the Riemann problem with data $(u^\flat, u^\sharp)$ may be solved without the introduction of two phase boundaries. Let $\mathcal{R}_1^\flat$ be the set of points that can be to the right of a wave of the first family exiting $u^\flat$. Similarly, let $\mathcal{L}_2^\sharp$ be the set of all those points that can be on the left of a wave of the second family entering $u^\sharp$. We require that

$$\inf_{u \in \mathcal{R}_1^\flat, w \in \mathcal{L}_2^\sharp} d\,(u, w) > \rho > 0\ . \tag{2.9}$$

We remark that without (2.9), the $\mathbf{L}^1$-continuous dependence on the initial data may be lost. Indeed, assume for simplicity that there exists a

$$u^* \in \mathcal{R}_1^\flat \cap \mathcal{L}_2^\sharp\ .$$

Choose now a positive a. Then, problem (1.1) with initial data

$$u_a(x) = \begin{cases} u^\flat & \text{if } x < -a, \\ u^\natural & \text{if } x \in [-a, a], \\ u^\sharp & \text{if } x > a \end{cases}$$

has a unique solution u_a containing two phase boundaries, due to (2.7) and (2.8). For all $a > 0$, the qualitative properties of the solution remain unchanged. However, due to Definition 2.1, at $a = 0$ the solution corresponding to u_0 contains no phase boundaries, but only a Lax wave of the first family joining $u^\flat$ to u^* and a Lax wave of the second family joining u^* to $u^\sharp$.

In case (2), a damping condition as in [5] is necessary to ensure that small perturbations of (1.2) still have a global solution. In order to state this condition, let us denote by $\Lambda^\flat$, $\Lambda^\sharp$ the propagation speeds of the phase boundaries, with $\Lambda^\flat, \Lambda^\sharp \in \mathbf{R}$. Consider the case of a small wave σ hitting one of the phase boundaries. From the interaction, a reflected wave and a transmitted wave arise. A first-order analysis shows that these arising waves are bounded by $|\sigma|$ times suitable reflection and transmission coefficients, respectively. These coefficients are given by

$$\Big(D_1\Phi r_1(u^\flat),\ -D_2\Phi r_2(u^\natural)\Big)^{-1}\Big(-D_2\Phi r_1(u^\flat),\ D_1\Phi r_2(u^\natural)\Big) = \begin{pmatrix} T_{\natural\flat} & R_\flat \\ R^\flat_\natural & T_{\flat\natural} \end{pmatrix}, \tag{2.10}$$

$$\Big(D_1\Phi r_1(u^\natural),\ -D_2\Phi r_2(u^\sharp)\Big)^{-1}\Big(-D_2\Phi r_1(u^\natural),\ D_1\Phi r_2(u^\sharp)\Big) = \begin{pmatrix} T_{\sharp\natural} & R^\sharp_\natural \\ R_\sharp & T_{\natural\sharp} \end{pmatrix}, \tag{2.11}$$

where for simplicity we omitted the arguments $(u^\flat, u^\natural)$ in the first line and $(u^\natural, u^\sharp)$ in the second one. Then let us define

$$\Theta^\flat = -\frac{\lambda_2(w^\natural) - \Lambda^\flat}{\Lambda^\flat - \lambda_1(w^\natural)}, \qquad \Theta^\sharp = -\frac{\lambda_1(w^\natural) - \Lambda^\sharp}{\Lambda^\sharp - \lambda_2(w^\natural)}.$$

Due to (2.4), both $\Theta^\flat$ and $\Theta^\sharp$ are negative numbers and, under condition (2.4), we have

$$\Theta^\flat \Theta^\sharp > 1. \tag{2.12}$$

We say that the *strong nonresonance* condition holds if

$$\left|R^\flat_\natural \Theta^\flat\right| \cdot \left|R^\sharp_\natural \Theta^\sharp\right| < 1 \tag{2.13}$$

with $R^\flat_\natural$ as in (2.10) and $R^\sharp_\natural$ as in (2.11). On one hand, this condition provides the Lipschitz-continuous dependence in $\mathbf{L}^1$ from the initial data of the solutions to (1.3); see [5]. On the other hand, it implies

$$\left|R^\flat_\natural\right| \cdot \left|R^\sharp_\natural\right| < 1 \tag{2.14}$$

because of (2.12). Formula (2.14) is the usual *nonresonance* condition (see [17], [5]) which says, roughly speaking, that the strength of a small wave diminishes after two reflections against the phase boundaries.

Let us remark that the stability conditions (2.7)–(2.8) together with the assumption

$$R^\flat_\natural \cdot R^\sharp_\natural \neq 1 \tag{2.15}$$

imply the stability of the solution to the Riemann problem (1.2) in case (iii) of Definition 2.1. In fact, by (2.10) and (2.11),

$$R^\flat_\natural = -\frac{\det T_1(u^\flat, u^\natural)}{\det S(u^\flat, u^\natural)}, \qquad R^\sharp_\natural = -\frac{\det T_2(u^\natural, u^\sharp)}{\det S(u^\natural, u^\sharp)},$$

where T_1, T_2, and S are the 2×2 matrices

$$\begin{aligned} T_i(u^l, u^r) &= \Big(D_1\Phi(u^l, u^r) r_i(u^l),\ D_2\Phi(u^l, u^r) r_i(u^r)\Big), \qquad i = 1, 2, \\ S(u^l, u^r) &= \Big(D_1\Phi(u^l, u^r) r_1(u^l),\ -D_2\Phi(u^l, u^r) r_2(u^r)\Big) . \end{aligned}$$

The stability of the solution to (1.2) in case (iii) of Definition 2.1 is equivalent to the applicability of the implicit function theorem to

$$\left\{ \begin{array}{l} \Phi\big(\phi_1(u^\flat, \sigma_1), u^\natural\big) = 0, \\ \Phi\big(u^\natural, \tilde{\phi}_2(u^\sharp, \sigma_2)\big) = 0 \end{array} \right.$$

in the unknowns σ_1, $u^\natural$, and σ_2. Here, ϕ_i (resp., $\tilde{\phi}_i$) is the shock-rarefaction curve from left to right (resp., right to left). The above requires the 4×4 matrix

$$\left(\begin{array}{ccc} D_1\Phi(u^\flat, u^\natural) r_1(u^\flat) & D_2\Phi(u^\flat, u^\natural) & 0 \\ 0 & D_1\Phi(u^\natural, u^\sharp) & D_2\Phi(u^\natural, u^\sharp) r_2(u^\sharp) \end{array} \right)$$

to be nonsingular, which in turn is equivalent to the nonsingularity of

$$\left(\begin{array}{cccc} D_1\Phi(u^\flat, u^\natural) r_1(u^\flat) & D_2\Phi(u^\flat, u^\natural) r_1(u^\natural) & D_2\Phi(u^\flat, u^\natural) r_2(u^\natural) & 0 \\ 0 & D_1\Phi(u^\natural, u^\sharp) r_1(u^\natural) & D_1\Phi(u^\natural, u^\sharp) r_2(u^\natural) & D_2\Phi(u^\natural, u^\sharp) r_2(u^\sharp) \end{array} \right).$$

A quick calculation shows that the determinant of the latter matrix is

$$\det T_1(u^\flat, u^\natural) \det T_2(u^\natural, u^\sharp) - \det S(u^\flat, u^\natural) \det S(u^\natural, u^\sharp),$$

which is nonzero iff (2.15) holds. This proves our remark.

The following definition of ΨRS is an adaptation of [2]; similar to [5], we introduce the set $\mathcal{M}$ of smooth increasing diffeomorphisms $\mathbf{R} \mapsto \mathbf{R}$. Assume that system (1.2) admits a solution $\underline{u}$, either in case (1) or in case (2).

Definition 2.2. *A Ψ-admissible Riemann semigroup (ΨRS) generated by* (1.2) *is a map $S\colon [0, +\infty[\times \mathcal{D} \mapsto \mathcal{D}$ satisfying the following:*

(i) *there exists a positive δ such that the closed invariant domain $\mathcal{D} \subset L^1_{loc}(\mathbf{R})$ contains the set of functions $u\colon \mathbf{R} \mapsto \Omega$, such that there exists $\mu \in \mathcal{M}$:*

$$\|u(\cdot) - \underline{u}(1, \mu(\cdot))\|_{\mathbf{L}^1} < \infty, \quad \mathrm{TV}\left\{u(\cdot) - \underline{u}(1, \mu(\cdot))\right\} \leq \delta; \tag{2.16}$$

(ii) $S_0 u = u$ *and* $S_{t''} \circ S_{t'}(u) = S_{t'+t''} u$;

(iii) *there exists a positive L such that*

$$\left\| S_{t''} u'' - S_{t'} u' \right\|_{\mathbf{L}^1} \leq L \cdot \left(\|u'' - u'\|_{\mathbf{L}^1} + |t'' - t'| \right);$$

(iv) *every trajectory $t \mapsto S_t u$ yields a Ψ-admissible weak solution to* (1.1) *with initial data u*;

(v) *if $u \in \mathcal{D}$ is piecewise constant, then for t small $S_t u$ coincides with the glueing of the Ψ-admissible solutions of Definition* 2.1.

Note that in case (2) the initial data u_o at (1.2) belongs to $\mathcal{D}$ but it does not belong to the set defined through (2.16). Moreover, the initial data satisfying (2.16) are more general than the ones considered at (1.3).

Our main result is the existence of such a ΨRS, which we accomplish by means of a *constructive* procedure.

Theorem 2.3. *Consider system* (1.1)*. Assume that it is strictly hyperbolic with each characteristic field either linearly degenerate or genuinely nonlinear. Let the admissibility condition* (2.5) *be given.*

In case (1)*, fix $u^\flat$ and $u^\sharp$ such that the solution to* (1.2) *contains a subsonic* (2.4) *and stable* (2.6) *phase boundary.*

In case (2)*, fix $u^\flat$, $u^\natural$, and $u^\sharp$ satisfying* (2.9) *and such that the solution to* (1.2) *contains two subsonic* (2.4)*, stable* (2.7)*–*(2.8)*, and strongly nonresonant* (2.13) *phase boundaries with middle state $u^\natural$.*

Then, there exists a ΨRS generated by (1.2)*.*

We remark here that the whole construction is *local* in the space of conserved quantities. Thus the above assumption on the linear degeneracy or genuine nonlinearity of the characteristic families is sufficient when satisfied in suitable neighborhoods of $u^\flat$, $u^\natural$, and $u^\sharp$.

3. Examples.

In this section we give two examples: the first coming from elastodynamics and the latter from van der Waals fluids.

In the case of elastodynamics we write $u = (v, w)$, where v is the particle velocity and w the strain; the flow function is $f(v, w) = (-\sigma(w), -v)$, defined on $\Omega = \mathbf{R} \times (]-1, w_M[\cup]w_m, +\infty[)$, with $w_M < w_m$. The system (1.1) becomes

$$\begin{cases} \partial_t v - \partial_x [\sigma(w)] = 0, \\ \partial_t w - \partial_x v = 0. \end{cases} \tag{3.1}$$

The function σ is the stress-strain function. We assume that it has a maximum point w_M and a minimum point w_m; see Figure 3.1. Moreover, with a little abuse of notation, we let

$$\Omega_1 =]-1, w_M[, \quad \Omega_2 =]w_m, +\infty[$$

and we assume also that

$$\sigma' > 0 \text{ in } \Omega_1 \cup \Omega_2 \qquad \text{and} \qquad \sigma'' < 0 \text{ in } \Omega_1, \quad \sigma'' > 0 \text{ in } \Omega_2; \tag{3.2}$$

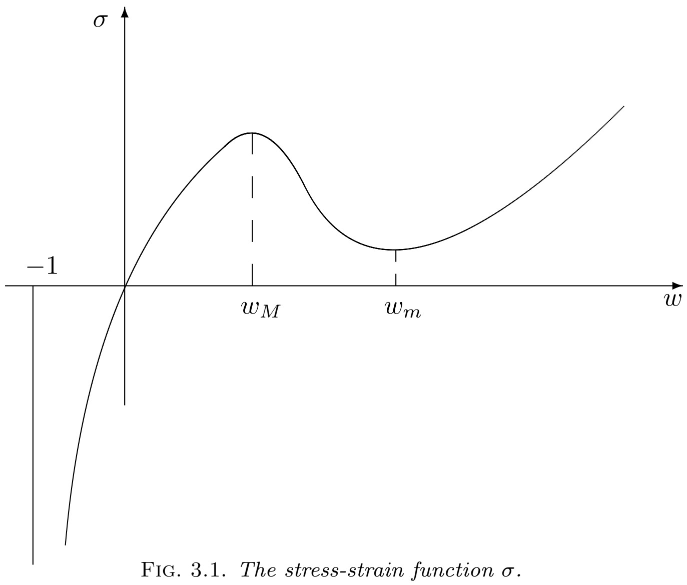


FIG. 3.1. *The stress-strain function* σ.

the behavior of σ'' in the middle zone $[w_M, w_m]$ is not relevant. In what follows, we shall always limit w to the phases Ω_1 or Ω_2, where the system is strictly hyperbolic and genuinely nonlinear. We introduce the sound speed

$$c(w) = \sqrt{\sigma'(w)}$$

so that the characteristic speeds of (3.1) are

$$\lambda_1(w) = -c(w) \quad \text{and} \quad \lambda_2(w) = c(w).$$

Let us remark that, from a physical point of view, both evolution from the first (hard) phase to the second (soft) one and vice versa are possible: an example of the first case is given by the cold drawing of polyethylene, while an example of the second case is the hardening of polybutene (see [23]). Moreover, the assumption of concavity of σ in the first phase and convexity in the second one is made only for simplicity; for some materials, different situations are possible and can be considered within the present construction.

We consider the Riemann problem for (3.1) with initial data

$$u_o(x) = \begin{cases} u^\flat = (v^\flat, w^\flat) & \text{if } x < 0, \\ u^\sharp = (v^\sharp, w^\sharp) & \text{if } x > 0. \end{cases} \tag{3.3}$$

Due to the nonmonotonicity of the stress-strain function σ, the Lax shock-rarefaction curves through $(v^\flat, w^\flat)$ and $(v^\sharp, w^\sharp)$ may have no intersection even if both $w^\flat$ and $w^\sharp$ both belong to the same hyperbolic phase. In fact, straightforward computations show that the Lax solution to (3.1) and (3.3) exists precisely in the following situations:

$$v^\sharp - v^\flat < \int_{w^\flat}^{w_M} c(w)dw + \int_{w^\sharp}^{w_M} c(w)dw \quad \text{if} \quad w^\flat, w^\sharp \in \Omega_1, \tag{3.4}$$

$$v^\flat - v^\sharp < \int_{w_m}^{w^\flat} c(w)dw + \int_{w_m}^{w^\sharp} c(w)dw \quad \text{if} \quad w^\flat, w^\sharp \in \Omega_2\ . \tag{3.5}$$

Condition (2.9) becomes

$$\begin{aligned} v^\sharp - v^\flat &> \int_{w^\flat}^{w_M} c(w)dw + \int_{w^\sharp}^{w_M} c(w)dw + \rho \quad \text{if} \quad w^\flat, w^\sharp \in \Omega_1, \\ v^\flat - v^\sharp &> \int_{w_m}^{w^\flat} c(w)dw + \int_{w_m}^{w^\sharp} c(w)dw + \rho \quad \text{if} \quad w^\flat, w^\sharp \in \Omega_2 \ . \end{aligned}$$

If $\int_{-1}^{w_M} c(w)dw < +\infty$, then the above inequalities follow from

$$\left\| u^\sharp - u^\flat \right\| > C \cdot \rho \ . \tag{3.6}$$

On the contrary, if $\int_{-1}^{w_M} c(w)dw = +\infty$, then no assumption of the type (3.6) implies (2.9).

Consider now a phase boundary with speed Λ, with right and left states $u^r = (v^r, w^r)$, $u^l = (v^l, w^l)$, respectively. From the Rankine–Hugoniot conditions (2.2), it follows that

$$\Lambda = \zeta \cdot \sqrt{\frac{\sigma(w^r) - \sigma(w^l)}{w^r - w^l}}, \tag{3.7}$$

where

$$\zeta = -\zeta_v \cdot \zeta_w, \quad \zeta_v = \operatorname{sign}(v^r - v^l), \quad \zeta_w = \operatorname{sign}(w^r - w^l) \tag{3.8}$$

(and $\operatorname{sign} 0 = 0$). We consider now admissibility and stability of the phase boundaries. In the kinetic approach proposed by [1] the choice for the function Ψ is

$$\begin{aligned} &\Psi(u^l, u^r) \\ &= \frac{\sigma(w^r) + \sigma(w^l)}{2}(w^r - w^l) - \int_{w^l}^{w^r} \sigma(w)\, dw + \zeta_w \phi \left(-\zeta_v \sqrt{\frac{\sigma(w^r) - \sigma(w^l)}{w^r - w^l}} \right), \end{aligned} \tag{3.9}$$

where ϕ is a given constitutive function and ζ_v, ζ_w are defined in (3.8).

If u^l, u^r are the side states of a nonstationary phase boundary, then the vanishing of the function Ψ prescribes the amount of physical entropy dissipated by the phase boundary.

The function ϕ above, as chosen in [1], is singular when $\Lambda = 0$. For this reason the case of a stationary phase boundary needs to be ruled out. With a different (smooth) choice of ϕ, the construction developed in section 2 can be applied also to the case of stationary phase boundaries.

In the case (3.9), the stability condition (2.6) reads (see [6])

$$\phi' \left(-\zeta_v \sqrt{\frac{\sigma(w^\sharp) - \sigma(w^\flat)}{w^\sharp - w^\flat}} \right) \neq - \frac{\dfrac{\sigma(w^\sharp) - \sigma(w^\flat)}{w^\sharp - w^\flat} + c(w^\flat)c(w^\sharp)}{c(w^\flat) + c(w^\sharp)} \, (w^\sharp - w^\flat)^2.$$

The above inequality is satisfied whenever ϕ is increasing, which is physically acceptable.

Let us consider now the case of two phase boundaries, having speeds $\Lambda^\flat$ and $\Lambda^\sharp$, with $\Lambda^\flat, \Lambda^\sharp \in \mathbf{R}$ and $\Lambda^\flat < \Lambda^\sharp$, with the same notations of Case 2 of the previous section; we look for an explicit condition for strong nonresonance. It comes out that

$$\text{(3.10)}\ R^\flat_\natural = -\frac{1}{\Theta^\flat}\frac{c(w^\natural)c(w^\flat) - (\Lambda^\flat)^2 + h^\flat\big(c(w^\natural) - c(w^\flat)\big)}{c(w^\natural)c(w^\flat) + (\Lambda^\flat)^2 + h^\flat\big(c(w^\natural) + c(w^\flat)\big)} \quad \text{for} \quad h^\flat = \frac{\phi'(\zeta^\flat_w \Lambda^\flat)}{(w^\natural - w^\flat)^2},$$

$$\text{(3.11)}\ R^\sharp_\natural = -\frac{1}{\Theta^\sharp}\frac{c(w^\natural)c(w^\sharp) - (\Lambda^\sharp)^2 + h^\sharp\big(c(w^\natural) - c(w^\sharp)\big)}{c(w^\natural)c(w^\sharp) + (\Lambda^\sharp)^2 + h^\sharp\big(c(w^\natural) + c(w^\sharp)\big)} \quad \text{for} \quad h^\sharp = \frac{\phi'(\zeta^\sharp_w \Lambda^\sharp)}{(w^\sharp - w^\natural)^2},$$

where

$$\Theta^\flat = \frac{\Lambda^\flat - c(w^\natural)}{\Lambda^\flat + c(w^\natural)}, \qquad \Theta^\sharp = \frac{\Lambda^\sharp + c(w^\natural)}{\Lambda^\sharp - c(w^\natural)}.$$

We already remarked in section 2 that $\Theta^\flat$ and $\Theta^\sharp$ are negative numbers and that (2.12) holds. Let us point out that, differently from [5], we have $\Theta^\flat < -1$ iff $\Lambda^\flat < 0$ and $\Theta^\sharp < -1$ iff $\Lambda^\sharp > 0$; the condition $\Lambda^\flat < 0 < \Lambda^\sharp$ is always satisfied in the case of a trilinear stress-strain function σ (see [1, p. 137]). What is more important here is that from (3.10) and (3.11) we see that the strong nonresonance condition (2.13) holds if the constitutive function ϕ is increasing.

We briefly note here that the approach developed in section 2 applies also to the case of those materials for which $\sigma'' < 0$ also in Ω_2. Indeed, in this case, simply substitute (3.5) with

$$v^\flat - v^\sharp < \sqrt{(\sigma(w^\sharp) - \sigma(w_m))\,(w^\sharp - w_m)} + \sqrt{\big(\sigma(w^\flat) - \sigma(w_m)\big)\,(w^\flat - w_m)};$$

here, $w^\flat$, $w^\sharp$ both belong to Ω_2. Similarly, condition (2.9) becomes

$$v^\flat - v^\sharp > \sqrt{(\sigma(w^\sharp) - \sigma(w_m))\,(w^\sharp - w_m)} + \sqrt{\big(\sigma(w^\flat) - \sigma(w_m)\big)\,(w^\flat - w_m)} + \rho.$$

The case $\sigma'' > 0$ in Ω_1 can be considered similarly.

With suitable choices of f, Ω_1, and Ω_2, the construction presented in this paper includes also the trilinear material, as considered in [13]. In this context, Definition 2.1 reduces to the one given therein. Moreover, see [21] for a simple cubic model of σ.

Our second example is the one-dimensional isothermal model for a van der Waals fluid:

$$\text{(3.12)} \qquad \begin{cases} \partial_t v + \partial_x\,[p(w)] = 0, \\ \partial_t w - \partial_x v = 0. \end{cases}$$

Here $u = (v, w)$, with v being the particle velocity and w the specific volume. The pressure p is a smooth positive function defined in $]0, +\infty[$, with a minimum point w_m and a maximum point $w_M > w_m$, $p(w_m) < p(w_M)$. The liquid phase is $\Omega_1 =]w_0, w_m[$ while the vapor phase is $\Omega_2 =]w_M, +\infty[$ and we assume

$$\text{(3.13)} \qquad p' < 0 \text{ in } \Omega_1 \cup \Omega_2, \qquad p'' > 0 \text{ in } \Omega_1,$$

as in Figure 3.2.

A key difference between this example and the previous one is the following. As shown in Figure 3.2, the pressure has an inflection point in Ω_2, which makes each of

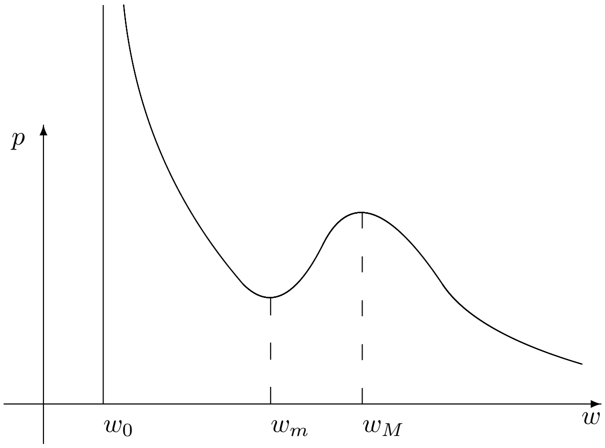


FIG. 3.2. *The pression p as a function of the specific volume w.*

the characteristic families neither globally linearly degenerate nor globally genuinely nonlinear. However, thanks to its *local* nature, the present construction still applies, provided that none of the points $u^\flat$, $u^\natural$, $u^\sharp$ coincides with the inflection point. We point out that in this case the sets $\mathcal{R}_1^\flat$ and $\mathcal{L}_2^\sharp$ in (2.9) are the *mixed curves* introduced by T.-P. Liu in [15, Section 2]. These curves generalize the usual Lax shock-rarefaction curves enabling the construction of a solution to Riemann problems with nongenuinely nonlinear characteristic families.

The formal calculations are omitted, since they are entirely analogous to what has been shown above. Let us emphasize, however, that the present construction can be applied also to the admissibility condition obtained in terms of the viscosity-capillarity criterium (see [18], [19], [8], [9], and [21, formula (5.26)] for the explicit admissibility condition) as well as to the so-called "normal growth" condition (see [22, formulas (3.6) and (4.11)]).

4. The algorithm. Following [4], all the construction below essentially relies on a suitable approximation of the solution to Riemann problems with data in Ω. As long as the solution to (1.2) does not develop any phase boundary, the results in [4] apply.

Consider now Cases 1 and 2. In Case 1, we construct two sets of Riemann coordinates, defined on neighborhoods $\mathcal{U}^\flat$, $\mathcal{U}^\sharp$ of $u^\flat$, $u^\sharp$. In Case 2, we introduce Riemann coordinates on disjoint neighborhoods $\mathcal{U}^\flat$, $\mathcal{U}^\natural$, and $\mathcal{U}^\sharp$ of $u^\flat$, $u^\natural$, and $u^\sharp$, which is possible due to (2.9). The local Riemann coordinates are denoted by v. Whenever necessary, the functions defined on $\mathcal{U}^\flat$, $(\mathcal{U}^\natural, \mathcal{U}^\sharp)$ will be bounded uniformly on $\mathcal{U}^\flat$, $(\mathcal{U}^\natural, \mathcal{U}^\sharp)$.

We define now the approximate solutions to the Riemann problems when the initial data belong to the previous neighborhoods.

In a given set of Riemann coordinates, the i-rarefaction curve ϕ_i^+ and the i-shock curve ϕ_i^- through point v can now be parametrized by means of the arc-length σ, for σ in a suitable neighborhood of 0, as

$$\begin{cases} \phi_1^+(v,\sigma) = (v_1+\sigma, v_2), \\ \phi_2^+(v,\sigma) = (v_1, v_2+\sigma), \end{cases} \qquad \begin{cases} \phi_1^-(v,\sigma) = (v_1+\sigma, v_2+\hat{\phi}_2(v,\sigma)\sigma^3), \\ \phi_2^-(v,\sigma) = (v_1+\hat{\phi}_1(v,\sigma)\sigma^3, v_2+\sigma) \end{cases}$$

for suitable smooth functions $\hat{\phi}_1, \hat{\phi}_2$. The juxtaposition of ϕ_i^+ and ϕ_i^- is the ith shock-rarefaction curve through v. Choose any $\mathbf{C}^\infty$ function $\varphi: \mathbf{R} \mapsto \mathbf{R}$ such that

$$\begin{cases} \varphi(s) = 1 & \text{if } s \le -2, \\ \varphi'(s) \in [-2, 0] & \text{if } s \in [-2, -1], \\ \varphi(s) = 0 & \text{if } s \ge -1, \end{cases}$$

and, for a fixed $\varepsilon > 0$, approximate the i-shock-rarefaction curve as follows:

$$\psi_i^\varepsilon(v, \sigma) = \varphi\left(\sigma/\sqrt{\varepsilon}\right) \cdot \phi_i^-(v, \sigma) + \left(1 - \varphi\left(\sigma/\sqrt{\varepsilon}\right)\right) \cdot \phi_i^+(v, \sigma), \quad i = 1, 2. \tag{4.1}$$

Let now a left and a right state u^l, u^r are given, with Riemann coordinates $v^l = (v_1^l, v_2^l)$ and $v^r = (v_1^r, v_2^r)$. There are two different situations.

A first possibility is that u^l, u^r both belong to the domain of the same chart (hence also to the same phase) and that the solution to the Riemann problem (1.1) with data

$$\bar{u}(x) = \begin{cases} u^l & \text{if } x < 0, \\ u^r & \text{if } x > 0 \end{cases} \tag{4.2}$$

attains values in the same phase. An ε-approximate solution to (1.1)–(4.2) is constructed as follows. First, by the implicit function theorem, we determine unique values σ_1 and σ_2 and a middle state v^m such that

$$v^r = \psi_2^\varepsilon\left(v^m, \sigma_2\right), \qquad v^m = \psi_1^\varepsilon\left(v^l, \sigma_1\right).$$

If $\sigma_1 \ge 0$, we connect the states v^l, v^m with a discrete rarefaction wave by the following procedure. Let the integers h, k be such that

$$h\varepsilon \le v_1^l < (h+1)\varepsilon, \qquad k\varepsilon \le v_1^m < (k+1)\varepsilon$$

and define the states

$$\omega_1^j = (j\varepsilon, v_2^l), \qquad \hat{\omega}_1^j = \left(\frac{2j+1}{2}\varepsilon, v_2^l\right) \qquad \text{for } j = h, \dots, k.$$

Then the ε-approximate solution in the quadrant $\{t \ge 0,\ x \le 0\}$ is the discrete rarefaction fan:

$$v^\varepsilon(t, x) = \begin{cases} v^l & \text{if } x < \lambda_1(\hat{\omega}_1^h)\, t, \\ \omega_1^j & \text{if } \lambda_1(\hat{\omega}_1^{j-1})\, t < x < \lambda_1(\hat{\omega}_1^j)\, t, \qquad j = h+1, \dots, k, \\ v^m & \text{if } \lambda_1(\hat{\omega}_1^k) < x \le 0. \end{cases} \tag{4.3}$$

If $\sigma_1 < 0$, the states v^l and v^m are connected by a single discontinuity:

$$v^\varepsilon(t, x) = \begin{cases} v^l & \text{if } x < \lambda_1^\varphi(v^l, \sigma_1)\, t, \\ v^m & \text{if } \lambda_1^\varphi(v^l, \sigma_1)\, t < x \le 0. \end{cases} \tag{4.4}$$

The speed λ_1^φ of the discontinuity is defined here as

$$\lambda_1^\varphi(v^l, \sigma_1) = \varphi(\sigma_1/\sqrt{\varepsilon}) \cdot \lambda_1^s(v^l, \sigma_1) + \left(1 - \varphi(\sigma_1/\sqrt{\varepsilon})\right) \cdot \lambda_1^r(v^l, \sigma_1),$$

with

$$\lambda_1^s(v^l, \sigma_1) = \lambda_1\left(v^l, \phi_1^-(v^l, \sigma_1)\right),$$

$$\lambda_1^r(v^l, \sigma_1) = \sum_{j=h}^{k} \frac{\operatorname{meas}\left([j\varepsilon, (j+1)\varepsilon] \cup [v_1^m, v_1^l]\right)}{|\sigma_1|}\ \lambda_1(\hat{\omega}_1^j)\ .$$

Observe that as soon as $\sigma_1 \leq -2\sqrt{\varepsilon}$ the function v^ε in (4.4) is an exact solution to the Rankine–Hugoniot equations, a shock wave. For this reason, we shall call briefly shock waves every funtion v^ε as defined in (4.4).

The construction of the ε-approximate solution on the quadrant where $x \geq 0$ is entirely similar, repeating the above construction with waves of the second family. We refer the reader to [4] for details.

A second eventuality is that u^l, u^r belong to different phases; in this case the Riemann problem is solved with the introduction of a small 1-wave, a *single exact* phase boundary, and a small 2-wave. This is possible by the implicit funtion theorem and the stability condition (2.6).

We pass now to a piecewise constant initial condition $\bar{u}$ belonging to some suitable domain. An ε-approximate piecewise constant solution to the Cauchy problem with initial data $\bar{u}$ is constructed as follows. At the initial time $\tau_0 = 0$ we solve the Riemann problems determined by the jumps in $\bar{u}$ applying the algorithm previously described. This yields a piecewise constant approximate solution $u = u^\varepsilon(t, x)$ defined up to the time $\tau_1 > 0$, where the first set of wave-front interactions takes place. We then solve these new Riemann problems by again applying the above algorithm. The solution is prolonged up to the time τ_2 where the second set of interactions takes place and so on.

The domain $\mathcal{D}^\varepsilon_\delta$ of the approximate semigroup in the two Cases 1 and 2 is defined as follows. We are concerned only with piecewise constant functions $u = u(x)$ of the form

$$u = u^\flat \cdot \chi_{]-\infty, x_1]} + \sum_{\alpha=1}^{n-1} u^\alpha \cdot \chi_{]x_\alpha, x_{\alpha+1}]} + u^\sharp \cdot \chi_{]x_n, +\infty[} \; . \tag{4.5}$$

Whenever $u^{\alpha-1}$, u^α belong to the same chart (and hence to the same phase), a suitable choice of neighborhoods $B\big(u^\flat, \delta_o\big) \subset \mathcal{U}^\flat$, $B\big(u^\natural, \delta_o\big) \subset \mathcal{U}^\natural$, and $B\big(u^\sharp, \delta_o\big) \subset \mathcal{U}^\sharp$ ensures that the Riemann problem determined by the jump at x_α is uniquely solved by the above algorithm in terms of waves with sizes $\sigma_{1,\alpha}$, $\sigma_{2,\alpha}$. Recalling (4.1), this means

$$v^\alpha = \psi_2^\varepsilon \left(\psi_1^\varepsilon(v^{\alpha-1}, \sigma_{1,\alpha}), \sigma_{2,\alpha} \right), \tag{4.6}$$

where $v^{\alpha-1}$, v^α stand for the Riemann coordinates of $u^{\alpha-1}$, u^α.

Case 1. Assume $\mathcal{U}^\flat \subseteq \Omega_1$ and $\mathcal{U}^\sharp \subseteq \Omega_2$. For all functions u of the form (4.5)–(4.6) satisfying

$$u^1, \ldots, u^{\alpha^\flat} \in B\Big(u^\flat, \delta_o\Big) \qquad \text{and} \qquad u^{\alpha^\flat+1}, \ldots, u^n \in B\big(u^\sharp, \delta_o\big), \tag{4.7}$$

define $\mathcal{A}^\flat$ as the set of pairs of waves $\sigma_{i,\alpha}$, $\sigma_{j,\beta}$, that are approaching (see [20]), with $\alpha, \beta \leq \alpha^\flat$; the set $\mathcal{A}^\sharp$ is defined similarly $(\alpha, \beta \geq \alpha^\flat + 1)$. Then we introduce the linear functionals and the interaction potentials as

$$\begin{aligned} V^\flat &= \sum_{\alpha=1}^{\alpha^\flat} |\sigma_{1,\alpha}| + K_2 \sum_{\alpha=\alpha^\flat+1}^{n} |\sigma_{2,\alpha}|, \qquad & V^\sharp &= K_1 \sum_{\alpha=1}^{\alpha^\flat} |\sigma_{1,\alpha}| + \sum_{\alpha=\alpha^\flat+1}^{n} |\sigma_{2,\alpha}|, \\ Q^\flat &= \sum_{\mathcal{A}^\flat} |\sigma_{i,\alpha}\sigma_{j,\beta}|, & Q^\sharp &= \sum_{\mathcal{A}^\sharp} |\sigma_{i,\alpha}\sigma_{j,\beta}| \end{aligned} \tag{4.8}$$

and finally

$$\Upsilon^\flat = V^\flat + Q^\flat, \quad \Upsilon^\sharp = V^\sharp + Q^\sharp, \tag{4.9}$$

$$\Upsilon = \Upsilon^\flat + \Upsilon^\sharp. \tag{4.10}$$

For the sake of simplicity, we omitted the dependence on u. Note the introduction of the weights K_i on the waves eventually impinging the phase boundary. Below it is proved that the domain

$$\mathcal{D}^{\varepsilon}_{\delta} = \{u \text{ as in (4.5)–(4.6)–(4.7)};\ \Upsilon(u) \leq \delta\} \tag{4.11}$$

is positively invariant with respect to system (1.1).

Case 2. Similar to above, assume $\mathcal{U}^{\flat} \subseteq \Omega_1$, $\mathcal{U}^{\natural} \subseteq \Omega_2$, and $\mathcal{U}^{\sharp} \subseteq \Omega_1$. Moreover, for all functions of the form (4.5)–(4.6) satisfying

$$\begin{aligned} &u^1, \ldots, u^{\alpha^{\flat}} \in B\big(u^{\flat}, \delta_o\big),\\ &u^{\alpha^{\flat}+1}, \ldots, u^{\alpha^{\sharp}-1} \in B\big(u^{\natural}, \delta_o\big),\\ &u^{\alpha^{\sharp}}, \ldots, u^{n} \in B\big(u^{\sharp}, \delta_o\big), \end{aligned} \tag{4.12}$$

with $\alpha^{\flat} + 1 < \alpha^{\sharp} - 1$ (so that u attains values in both phases), introduce

$$\begin{aligned} &V^{\flat} = \sum_{i=1}^{2} \sum_{\alpha=1}^{\alpha^{\flat}} K_i^{\flat} |\sigma_{i,\alpha}|, \quad V^{\natural} = \sum_{i=1}^{2} \sum_{\alpha=\alpha^{\flat}+1}^{\alpha^{\sharp}-1} K_i^{\natural} |\sigma_{i,\alpha}|, \quad V^{\sharp} = \sum_{i=1}^{2} \sum_{\alpha=\alpha^{\sharp}}^{n} K_i^{\sharp} |\sigma_{i,\alpha}|,\\ &Q^{\flat} = \sum_{\mathcal{A}^{\flat}} |\sigma_{i,\alpha}\sigma_{j,\beta}|, \qquad Q^{\natural} = \sum_{\mathcal{A}^{\natural}} |\sigma_{i,\alpha}\sigma_{j,\beta}|, \qquad Q^{\sharp} = \sum_{\mathcal{A}^{\sharp}} |\sigma_{i,\alpha}\sigma_{j,\beta}| \end{aligned} \tag{4.13}$$

and then

$$\Upsilon^{\flat} = V^{\flat} + Q^{\flat}, \qquad \Upsilon^{\natural} = V^{\natural} + Q^{\natural}, \qquad \Upsilon^{\sharp} = V^{\sharp} + Q^{\sharp},$$

$$\Upsilon = \Upsilon^{\flat} + \Upsilon^{\natural} + \Upsilon^{\sharp} + \frac{1}{K_3} \left\| v^{\alpha^{\flat}+1} - v^{\natural} \right\|. \tag{4.14}$$

At last we define

$$\mathcal{D}^{\varepsilon}_{\delta} = \{u \text{ as in (4.5)–(4.6)–(4.12)};\ \Upsilon(u) \leq \delta\}\,. \tag{4.15}$$

The various constants K_i, $K_i^{\flat}$, $K_i^{\natural}$, $K_i^{\sharp}$ in (4.8), (4.13), and (4.14) will be defined later. They all depend on f and Φ_S.

PROPOSITION 4.1. *Let the Riemann problem* (1.2) *satisfy the stability assumption* (2.6) *in Case* 1, *the stability assumptions* (2.7), (2.8), *and the nonresonance condition* (2.14) *in Case* 2. *Then there exists* $\delta > 0$, *and suitable constants in the definitions* (4.8), (4.13), *and* (4.14), *independent of* ε *such that, for any* $\bar{u} \in \mathcal{D}^{\varepsilon}_{\delta}$, *the wave-front tracking algorithm constructs a unique approximate solution* $u^{\varepsilon} \colon [0, +\infty[\times \mathbf{R} \mapsto \mathbf{R}^2$ *of*

$$\begin{cases} \partial_t u + \partial_x \left[f(u)\right] = 0,\\ u(0,x) = \bar{u}(x) \end{cases} \tag{4.16}$$

with the following properties:

(i) $u^{\varepsilon}(t,\cdot) \in \mathcal{D}^{\varepsilon}_{\delta}$ *for all* $t \geq 0$;

(ii) *the function* $t \mapsto \Upsilon\left(u^{\varepsilon}(t,\cdot)\right)$ *is nonincreasing*;

(iii) *any strip* $[0,T] \times \mathbf{R}$ *contains finitely many interaction points of* u^{ε};

(iv) $\mathrm{TV}\left(u^{\varepsilon}(t,\cdot)\right)$ *is uniformly bounded*;

(v) u^{ε} *is* Ψ*-admissible.*

Here $\mathcal{D}^\varepsilon_\delta$ is defined by (4.11) *in Case* 1 *and by* (4.15) *in Case* 2*.*

In both cases, to denote this unique, globally defined, ε-approximate solution, we use the semigroup notation

$$u^\varepsilon(t,\cdot) = S^\varepsilon_t \bar{u}. \tag{4.17}$$

As in [4], the rest of the proof works toward an estimate independent from ε of the Lipschitz constant for S^ε, in the $\mathbf{L}^1$ norm. The basic technique is to shift the locations of the jumps of the initial data $\bar{u}$ at constant rates and then to estimate the shift rates of the jumps of the solution $u^\varepsilon(t,\cdot)$, at any fixed $t > 0$. First we introduce the shifts and the notion of pseudopolygonal.

Definition 4.2. *Let $]a,b[$ be an open interval. An elementary path is a map $\gamma\colon]a,b[\mapsto \mathbf{L}^1_{loc}(\mathbf{R})$ of the form*

$$\gamma(\theta) = \sum_{\alpha=1}^{N} u^\alpha \cdot \chi_{]x^\theta_{\alpha-1},\, x^\theta_\alpha[}, \qquad x^\theta_\alpha = x_\alpha + \xi_\alpha \theta, \tag{4.18}$$

with $x^\theta_{\alpha-1} < x^\theta_\alpha$ for all $\theta \in]a,b[$ and $\alpha = 1,\ldots,N$; the constants ξ_α are called shift rates.

Definition 4.3. *A continuous map $\gamma\colon [a,b] \mapsto \mathbf{L}^1_{loc}(\mathbf{R})$is called a pseudopolygonal if there exist countably many disjoint open intervals $J_h \subset [a,b]$ such that*

(i) *the restriction of γ to each J_h is an elementary path;*

(ii) *the set $[a,b] \setminus \bigcup_{h\geq 1} J_h$ is countable.*

Exactly as in [4], one can prove the following proposition.

Proposition 4.4. *Let $\gamma_o\colon [a,b] \mapsto \mathcal{D}^\varepsilon_\delta$ be a pseudopolygonal. Then, for all $\tau > 0$, the path*

$$\gamma_\tau = S^\varepsilon_\tau \circ \gamma_o$$

is also a pseudopolygonal. Indeed, there exist countably many open intervals J_h such that $[a,b] \setminus \bigcup J_h$ is countable and the wave-front configuration of the solution $u^\theta = S^\varepsilon_\tau \circ \gamma_o(\theta)$ on $[0,\tau] \times \mathbf{R}$ remains the same as θ ranges in each J_h.

Below, we move towards a definition of length of pseudopolygonals by first defining the length of elementary paths. The latter, in turn, depends on a suitable functional Υ_ξ which we define below.

Case 1. For all u in $\mathcal{D}^\varepsilon_\delta$ defined as in (4.11), let

$$\begin{aligned} V^\flat_\xi &= \sum_{i=1}^{2} \sum_{\alpha=1}^{\alpha^\flat} p^\flat_{i,\alpha} |\sigma_{i,\alpha} \xi_{i,\alpha}|, & Q^\flat_\xi &= \sum_{\mathcal{A}^\flat} |\sigma_{i,\alpha}\sigma_{j,\beta}| \left(p^\flat_{i,\alpha} |\xi_{i,\alpha}| + p^\flat_{j,\beta} |\xi_{j,\beta}| \right), \\ V^\sharp_\xi &= \sum_{i=1}^{2} \sum_{\alpha=\alpha^\flat+1}^{n} p^\sharp_{i,\alpha} |\sigma_{i,\alpha} \xi_{i,\alpha}|, & Q^\sharp_\xi &= \sum_{\mathcal{A}^\sharp} |\sigma_{i,\alpha}\sigma_{j,\beta}| \left(p^\sharp_{i,\alpha} |\xi_{i,\alpha}| + p^\sharp_{j,\beta} |\xi_{j,\beta}| \right). \end{aligned} \tag{4.19}$$

Then we define

$$\Upsilon^\flat_\xi = V^\flat_\xi (1 + Q^\flat) + K^\flat Q^\flat_\xi, \qquad \Upsilon^\sharp_\xi = V^\sharp_\xi (1 + Q^\sharp) + K^\sharp Q^\sharp_\xi \tag{4.20}$$

with $Q^\flat$ and $Q^\sharp$ defined as in (4.8) and finally

$$\Upsilon_\xi = \left(\Upsilon^\flat_\xi + \Upsilon^\sharp_\xi + \left|\hat{\xi}\right| \right) e^{\mathcal{K}\Upsilon}. \tag{4.21}$$

Above, $\hat{\xi}$ is the shift speed of the phase boundary. The constants $p^\flat_{i,\alpha}$, $p^\sharp_{i,\alpha}$, $K^\flat$, $K^\sharp$, and $\mathcal{K}$ are specified in section 5.

Case 2. For all u in $\mathcal{D}^\varepsilon_\delta$ as defined in (4.15), let

$$\begin{aligned} V^\flat_\xi &= \sum_{i=1}^{2} \sum_{\alpha=1}^{\alpha^\flat} p^\flat_{i,\alpha} |\sigma_{i,\alpha} \xi_{i,\alpha}|, & Q^\flat_\xi &= \sum_{\mathcal{A}^\flat} |\sigma_{i,\alpha}\sigma_{j,\beta}| \left(p^\flat_{i,\alpha}|\xi_{i,\alpha}| + p^\flat_{j,\beta}|\xi_{j,\beta}|\right), \\ V^\natural_\xi &= \sum_{i=1}^{2} \sum_{\alpha=\alpha^\flat+1}^{\alpha^\sharp-1} p^\natural_{i,\alpha} |\sigma_{i,\alpha} \xi_{i,\alpha}|, & Q^\natural_\xi &= \sum_{\mathcal{A}^\natural} |\sigma_{i,\alpha}\sigma_{j,\beta}| \left(p^\natural_{i,\alpha}|\xi_{i,\alpha}| + p^\natural_{j,\beta}|\xi_{j,\beta}|\right), \\ V^\sharp_\xi &= \sum_{i=1}^{2} \sum_{\alpha=\alpha^\sharp+1}^{n} p^\sharp_{i,\alpha} |\sigma_{i,\alpha} \xi_{i,\alpha}|, & Q^\sharp_\xi &= \sum_{\mathcal{A}^\sharp} |\sigma_{i,\alpha}\sigma_{j,\beta}| \left(p^\sharp_{i,\alpha}|\xi_{i,\alpha}| + p^\sharp_{j,\beta}|\xi_{j,\beta}|\right); \end{aligned} \tag{4.22}$$

then

$$\begin{aligned} \Upsilon^\flat_\xi &= V^\flat_\xi (1 + Q^\flat) + K^\flat Q^\flat_\xi, \\ \Upsilon^\natural_\xi &= V^\natural_\xi (1 + Q^\natural) + K^\natural Q^\natural_\xi, \\ \Upsilon^\sharp_\xi &= V^\sharp_\xi (1 + Q^\sharp) + K^\sharp Q^\sharp_\xi, \end{aligned} \tag{4.23}$$

and, finally,

$$\Upsilon_\xi = \left(\Upsilon^\flat_\xi + \Upsilon^\natural_\xi + \Upsilon^\sharp_\xi + H\left(\left|\hat{\xi}^\flat\right| + \left|\hat{\xi}^\sharp\right|\right)\right) e^{\mathcal{K}\Upsilon}; \tag{4.24}$$

$\hat{\xi}^\flat$ and $\hat{\xi}^\sharp$ are the shift speeds of the phase boundaries. Also in this case the constants $p^\flat_{i,\alpha}$, $p^\natural_{i,\alpha}$, $p^\sharp_{i,\alpha}$, $K^\flat$, $K^\natural$, $K^\sharp$, H, and $\mathcal{K}$ are given in section 5.

By means of Υ_ξ we can now define the weighted length of a polygonal and the weighted distance between two piecewise constant functions. Note that if γ is an elementary path, then the function $\theta \mapsto \Upsilon_\xi(\gamma(\theta))$ is constant.

DEFINITION 4.5. *For a fixed $\varepsilon > 0$ the weighted length of the elementary path γ in* (4.18) *is*

$$\|\gamma\| \doteq (b - a) \cdot \Upsilon_\xi(\gamma).$$

DEFINITION 4.6. *The weighted length of a pseudopolygonal is the sum of the weighted lengths of its elementary paths. For any two piecewise constant functions $u, w \in \mathcal{D}^\varepsilon_\delta$, their weighted distance is*

$$d_\varepsilon(u, w) \doteq \inf\left\{\|\gamma\|;\ \gamma\colon [0,1] \mapsto \mathcal{D}^\varepsilon_\delta \textit{ is a pseudopolygonal joining } u \textit{ with } w\right\}. \tag{4.25}$$

Below, we prove that the function

$$t \mapsto d_\varepsilon\left(S^\varepsilon_t u, S^\varepsilon_t u\right)$$

is nonincreasing for all $u, w \in \mathcal{D}^\varepsilon_\delta$. This, together with the equivalence of d_ε with the $\mathbf{L}^1$ distance, implies that the semigroup S^ε is uniformly Lipschitz continuous with respect to the $\mathbf{L}^1$ distance.

PROPOSITION 4.7. *Let the Riemann problem* (1.2) *satisfy the assumptions of Theorem* 2.3. *In both Cases* 1 *and* 2, *there exist $\delta > 0$ and positive constants in the above definitions of Υ_ξ, independent of ε, such that if γ_o is a pseudopolygonal,* *then the weighted length* $\|\gamma_\tau\|$ *of the pseudopolygonal* $\gamma_\tau = S^\varepsilon_\tau \circ \gamma_o$ *is a nonincreasing function of time.*

PROPOSITION 4.8. *For any* $\delta > 0$, *there exists some* $\delta' \in \,]0, \delta]$ *such that any two functions* u, u' *in* $\mathcal{D}^\varepsilon_{\delta'}$ *can be joined by a pseudopolygonal entirely contained in* $\mathcal{D}^\varepsilon_\delta$. *Moreover, the weighted length of this pseudopolygonal is uniformly equivalent with respect to* ε *to the usual distance* $\|u - u'\|_{\mathbf{L}^1}$.

PROPOSITION 4.9. *Let the Riemann problem* (1.2) *satisfy the assumptions of Theorem* 2.3. *Then there exists a positive* δ, *independent of* ε, *such that the semigroup*

$$S^\varepsilon\colon [0, +\infty[\, \times \mathcal{D}^\varepsilon_\delta \mapsto \mathcal{D}^\varepsilon_\delta$$

defined by (4.17) *is uniformly Lipschitz continuous with respect to the* $\mathbf{L}^1$ *distance, with a Lipschitz constant independent of* ε.

As in [4], to complete the proof of Theorem 2.3, we now consider a sequence of semigroups S^{ε_n} with $\lim_{n\to+\infty} \varepsilon_n = 0$ and construct the limit semigroup. More precisely, we fix $\delta > 0$ according to Proposition 4.9 and define the closed domain

$$\mathcal{D} = \{\bar{u}\colon \exists \bar{u}_n \to \bar{u}, \quad \bar{u}_n \in \mathcal{D}^\varepsilon_\delta \quad \text{for all } n\}. \tag{4.26}$$

For $\bar{u} \in \mathcal{D}$ and $t \geq 0$, we then define

$$S_t \bar{u} = \lim_{n\to+\infty} S^{\varepsilon_n}_t \bar{u}_n, \tag{4.27}$$

where $\bar{u}_n \in \mathcal{D}^{\varepsilon_n}_\delta$ is any sequence approaching $\bar{u}$ in $\mathbf{L}^1$. We conclude by proving the following proposition.

PROPOSITION 4.10. *The closed domain* $\mathcal{D}$ *in* (4.26) *and the semigroup* S *in* (4.27) *are well defined and satisfy* (i)–(v) *of Definition* 2.2 *for suitable constants* $L, \delta > 0$.

Let us point out that an important point in the previous proposition consists in proving that the Ψ-admissibility is conserved in the limit.

5. Technical proofs.

In this section we collect those technical parts that differ significantly from [4]. Indeed, the present construction of the semigroup differs from the one therein in the proofs that the amounts Υ and Υ_ξ are nonincreasing when an interaction involving a phase boundary takes place (Propositions 4.1, part (ii), and 4.7). Once this is known, the same inductive (resp., perturbation) technique used in [4] to prove that Υ (resp., Υ_ξ) is nonincreasing still applies. Then we give a proof of Proposition 4.8 on the basis of an analogous result in [5]. For what concerns Proposition 4.10, we prove the Ψ-admissibility of the orbits of the semigroup and refer the reader to [4] for the missing parts.

Propositions 4.4 and 4.9 are proved as in [4].

Aiming at the proof of Proposition 4.1 we recall the basic interaction estimates.

First, we give the basic estimates for two small interacting waves, the case of many interacting waves is covered as in [4, Lemma 5]. Here and in all that follows, σ_i^+ denotes the total size of outgoing i-waves; see (4.3).

With reference to the notation in Figure 5.1(i), if two waves σ_1^-, σ_2^- of different families interact, then

$$\left|\sigma_1^+ - \sigma_1^-\right| + \left|\sigma_2^+ - \sigma_2^-\right| \leq C \cdot \left|\sigma_1^- \sigma_2^-\right| \left(\left|\sigma_1^-\right| + \left|\sigma_2^-\right|\right). \tag{5.1}$$

In this section, by C we denote a suitably large positive constant. In the case of two waves σ', σ'' both belonging to the first family (see Figure 5.1(ii)), we have

$$\left|\sigma_1^+ - (\sigma' + \sigma'')\right| + \left|\sigma_2^+\right| \leq C \cdot |\sigma'\sigma''| \left(|\sigma'| + |\sigma''|\right). \tag{5.2}$$

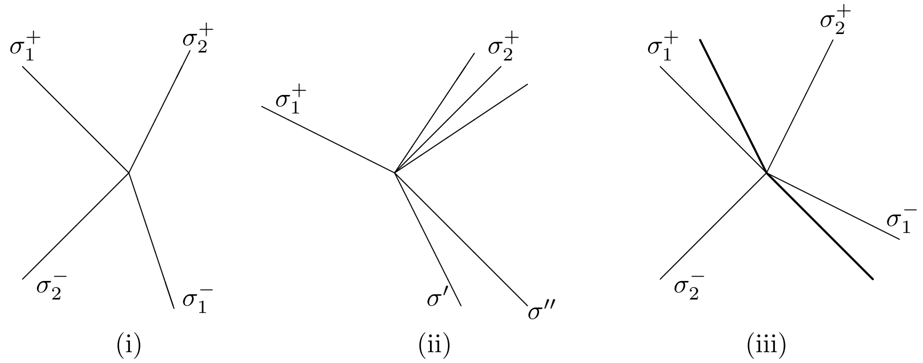


FIG. 5.1. *Interaction of small waves of different families* (i)*, of the same family* (ii)*, and of small waves with the phase boundary* (iii)*.*

The case of waves both belonging to the second family is entirely similar.

Moreover, we shall consider in detail interactions involving a phase boundary having, say, a left state in Ω_1 and a right state in Ω_2, with the notation as in Figure 5.1(iii).

In Case 1, using the notation in Figure 5.1(iii), the basic estimates related to interactions involving the phase boundary are

$$|\sigma_1^+| \le C\left(|\sigma_1^-| + |\sigma_2^-|\right), \qquad |\sigma_2^+| \le C\left(|\sigma_1^-| + |\sigma_2^-|\right). \tag{5.3}$$

In Case 2, slightly more precise estimates are necessary:

$$|\sigma_1^+| \le C\left(|\sigma_1^-| + |\sigma_2^-|\right), \qquad |\sigma_2^+| \le (1 + C\delta)\left|R_\natural^\flat\right||\sigma_1^-| + C|\sigma_2^-|. \tag{5.4}$$

An entirely similar convention is followed in case (2.11).

Let τ be a time at which an interaction occurs. We use the notation

$$F^+ = F(\tau+), \qquad F^- = F(\tau-), \qquad \text{and} \qquad \Delta F = F^+ - F^-,$$

where F can be V, Q, Υ, V_ξ, Q_ξ, or Υ_ξ. In general, the signs $+$ (resp., $-$) are attached to quantities related to waves exiting (resp., entering) an interaction.

Proof of Proposition 4.1. Below, we prove that Υ decreases strictly whenever a simple interaction takes place, in both Cases 1 and 2.

Case 1. We choose first

$$K_1 = K_2 = 1 + 2C$$

and a sufficiently small δ so that (5.5) and (5.6) hold.

(1) Two small waves of different families interact in the left phase.

We denote the waves as in Figure 5.1(i); by (5.1) we have

$$\begin{aligned}\Delta\Upsilon = \Delta\Upsilon^\flat &\le |\sigma_1^+| + K_2|\sigma_2^+| - |\sigma_1^-| - K_2|\sigma_2^-| - |\sigma_1^-\sigma_2^-| \\ &\qquad\qquad + \left(|\sigma_1^+ - \sigma_1^-| + |\sigma_2^+ - \sigma_2^-|\right)V^\flat \\ &\le \left[C(1 + K_2 + \delta)\left(|\sigma_1^-| + |\sigma_2^-|\right) - 1\right]\cdot|\sigma_1^-\sigma_2^-| \\ &< 0\end{aligned} \tag{5.5}$$

provided δ is sufficiently small.

(2) Two small waves of a same family interact in the left phase.

As in Figure 5.1(ii), σ' and σ'' belong to the first family (the other case is analogous). Then, by (5.2)

$$\begin{aligned}\Delta\Upsilon = \Delta\Upsilon^\flat &\le \left|\sigma_1^+\right| + K_2\left|\sigma_2^+\right| - |\sigma'| - |\sigma''| - |\sigma'\sigma''| + \left(\left|\sigma_1^+ - (\sigma' + \sigma'')\right| + \left|\sigma_2^+\right|\right) V^\flat \\ &\le \left[C\left(1 + K_2 + \delta\right)\left(|\sigma'| + |\sigma''|\right) - 1\right] \cdot |\sigma'\sigma''| \qquad (5.6) \\ &< 0\end{aligned}$$

provided δ is sufficiently small. The interaction of small waves in the right phase is completely analogous; hence it is omitted.

(3) Two small waves interact with the phase boundary.

The following estimates rely on a first-order argument. Thus we can consider interactions of, possibly, several waves hitting the phase boundary on both sides. Call σ_i^- the total size of the waves of the ith family impinging the phase boundary; then

$$\begin{aligned}\Delta\Upsilon &\le \left|\sigma_1^+\right| + \left|\sigma_2^+\right| + \left|\sigma_1^+\right|V^\flat + \left|\sigma_2^+\right|V^\sharp - K_2\left|\sigma_2^-\right| - K_1\left|\sigma_1^-\right| \\ &\le \left[C(1+\delta) - K_1\right]\left|\sigma_1^-\right| + \left[C(1+\delta) - K_2\right]\left|\sigma_2^-\right| \\ &\le -\left(\left|\sigma_1^-\right| + \left|\sigma_2^-\right|\right) \qquad (5.7)\end{aligned}$$

due to the choice of K_i. The proof of Proposition 4.1 in Case 1 is concluded.

Case 2. We choose the weights in (4.13) in the following way. First

$$K_1^\flat = 1 \quad \text{and} \quad K_2^\sharp = 1.$$

Because of (2.14) we can choose $K_1^\natural$ and $K_2^\natural$ such that

$$\begin{aligned}&K_1^\natural - (1 + C\delta)\left|R_\natural^\flat\right|K_2^\natural > 2C + (1 + C\delta)\left|R_\natural^\flat\right| + 2, \\ &K_2^\natural - (1 + C\delta)\left|R_\natural^\sharp\right|K_1^\natural > 2C + (1 + C\delta)\left|R_\natural^\sharp\right| + 2\ .\end{aligned} \qquad (5.8)$$

Next select $K_2^\flat$, $K_1^\sharp$ sufficiently large so that estimates analogous to the ones in Case 1, subcase 3, still hold and moreover

$$K_2^\flat > C(K_2^\natural + 3) + 2 \quad \text{and} \quad K_1^\sharp > C(K_1^\natural + 3) + 2\,. \qquad (5.9)$$

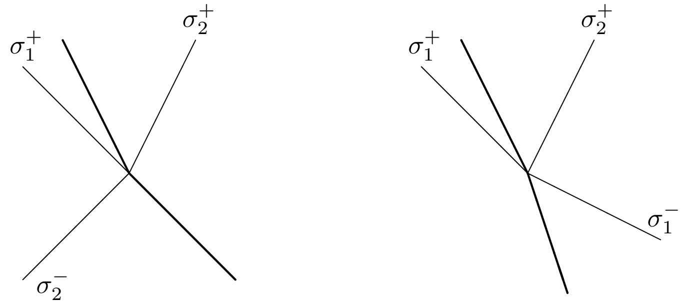
Let small waves of strengths σ_1^- and σ_2^- impinge on the leftmost phase boundary; see Figure 5.1(iii). Let v_r^-, v_r^+ be the states in $\mathcal{U}^\natural$ just on the right of the phase boundary, respectively, before and after the interaction. A simple first-order argument shows that there exists a constant C such that

$$\|v_r^+ - v_r^-\| \le C\left(\left|\sigma_1^-\right| + \left|\sigma_2^-\right|\right); \qquad (5.10)$$

let K_3 in (4.14) be such that $K_3 \ge C$.

We consider in detail only the case of Figure 5.1(iii), since the interactions against the other phase boundary are treated entirely similarly, while those interactions far from the phase boundaries can be tackled as in Case 1. From (5.4) we have

$$\begin{aligned}\Delta\Upsilon^\flat &\le K_1^\flat\left|\sigma_1^+\right| - K_2^\flat\left|\sigma_2^-\right| + \left|\sigma_1^+\right|V^{\flat-} \\ &\le \left(C(K_1^\flat + \delta) - K_2^\flat\right)\left|\sigma_2^-\right| + C(K_1^\flat + \delta)\left|\sigma_1^-\right|, \\ \Delta\Upsilon^\natural &\le K_2^\natural\left|\sigma_2^+\right| - K_1^\natural\left|\sigma_1^-\right| + \left|\sigma_2^+\right|V^{\natural-} \\ &\le \left((1 + C\delta)\left|R_\natural^\flat\right|(K_2^\natural + \delta) - K_1^\natural\right)\left|\sigma_1^-\right| + C(K_2^\natural + \delta)\left|\sigma_2^-\right|.\end{aligned}$$



(i) (ii)

FIG. 5.2. *Interactions of waves (thin lines) with the phase boundary (thick line).*

Then by (5.8), (5.9), and (5.10) we deduce

$$\begin{aligned}\Delta\Upsilon &\le \Big((1+C\delta)\Big|R^\flat_\natural\Big|(K_2^\natural+\delta)+C(K_1^\flat+\delta)-K_1^\natural\Big)\left|\sigma_1^-\right| \\ &\quad+\Big(C(K_1^\flat+\delta)+C(K_2^\natural+\delta)-K_2^\flat\Big)\left|\sigma_2^-\right| \\ &\quad+\left(\left|\sigma_1^-\right|+\left|\sigma_2^-\right|\right) \\ &\le -\left(\left|\sigma_1^-\right|+\left|\sigma_2^-\right|\right) \qquad (5.11)\end{aligned}$$

due to (5.9) and to the choice of K_3. This accounts for Case 2; the proof of Proposition 4.1 is therefore complete. □

We note that applying the usual compactness argument based on Helly's theorem to the result of the above Proposition 4.1 we have the usual Glimm's theorem on the global existence of weak solutions to (1.1).

Preliminarily to the proof Proposition 4.7, we collect below the basic interaction estimates for shifting waves; see [4]. It goes without saying that ξ_i^+ (resp., ξ_i^-, ξ', etc.) are the shift speeds of σ_i^+ (resp., σ_i^-, σ', etc.); see Figures 5.1(i) and 5.1(ii).

$$\begin{aligned}&\sum_\alpha\left|\sigma_{1,\alpha}^+\xi_{1,\alpha}^+\right|-\left|\sigma_1^-\xi_1^-\right|+\sum_\alpha\left|\sigma_{2,\alpha}^+\xi_{2,\alpha}^+\right|-\left|\sigma_2^-\xi_2^-\right| \\ &\qquad\le C\cdot\left|\sigma_1^-\sigma_2^-\right|\cdot\left(\left|\xi_1^-\right|+\left|\xi_2^-\right|\right), \qquad (5.12)\end{aligned}$$

$$\left|\sigma_1^+\xi_1^+\right|-\left(\left|\sigma' xi'\right|+\left|\sigma'' xi''\right|\right)+\sum_\alpha\left|\sigma_{2,\alpha}^+\xi_{2,\alpha}^+\right|\le C\cdot\left|\sigma'\sigma''\right|\cdot\left(\left|\xi'\right|+\left|\xi''\right|\right). \qquad (5.13)$$

We refer to [4, Lemmas 21, 22] for a proof.

In Case 1, refer to Figures 5.2(i) and 5.2(ii) and denote by $\hat\xi^-$, $\hat\xi^+$ the shift speeds of the phase boundaries before and after the interaction.

A first-order analysis similar to the one in [5] yields

$$\left|\hat\xi^+-\hat\xi^-\right|\le C\cdot\left|\sigma_i^-\right|\left(\left|\xi_i^-\right|+\left|\hat\xi^-\right|\right), \qquad (5.14)$$

$$\sum_\alpha\left|\sigma_{1,\alpha}^+\xi_{1,\alpha}^+\right|\le C\cdot\left|\sigma_i^-\right|\left(\left|\xi_i^-\right|+\left|\hat\xi^-\right|\right), \qquad (5.15)$$

$$\sum_\alpha\left|\sigma_{2,\alpha}^+\xi_{2,\alpha}^+\right|\le C\cdot\left|\sigma_i^-\right|\left(\left|\xi_i^-\right|+\left|\hat\xi^-\right|\right) \qquad (5.16)$$

for $i=1,2$.

In Case 2, we make use of the same estimates, but we denote the shift speeds of the two phase boundaries by $\xi^\flat$ and $\xi^\sharp$. Moreover, (see (5.20) in [5])

$$\sum_\alpha \left|\sigma_{2,\alpha}^+ \xi_{2,\alpha}^+\right| \leq (1+C\delta)\left|R^\flat_\natural \Theta^\flat\right| \left|\sigma_1^- \xi_1^-\right| + C\cdot\left|\sigma_1^-\right|\left|\xi^{\flat -}\right|, \tag{5.17}$$

$$\sum_\alpha \left|\sigma_{1,\alpha}^+ \xi_{1,\alpha}^+\right| \leq (1+C\delta)\left|R^\sharp_\natural \Theta^\sharp\right| \left|\sigma_2^- \xi_2^-\right| + C\cdot\left|\sigma_2^-\right|\left|\xi^{\sharp -}\right|. \tag{5.18}$$

Proof of Proposition 4.7. We denote the quantities related to interacting waves as above; see Figures 5.1(i), 5.1(ii), 5.1(iii), 5.2(i), and 5.2(ii). Noninteracting waves are labeled $\sigma_{j,\beta}$. We denote, moreover, by $\mathcal{A}_{i,\alpha}$ the set of waves approaching $\sigma_{i,\alpha}$.

Case 1. First, the choice of the weights; referring to formula (4.19) we take

$$\begin{aligned} p^\flat_{1,\alpha} &= 1 + \varepsilon_o \operatorname{sign} \sigma_{1,\alpha}, & p^\flat_{2,\alpha} &= K(1 + \varepsilon_o \operatorname{sign} \sigma_{2,\alpha}), \\ p^\sharp_{1,\alpha} &= K(1 + \varepsilon_o \operatorname{sign} \sigma_{1,\alpha}), & p^\sharp_{2,\alpha} &= 1 + \varepsilon_o \operatorname{sign} \sigma_{2,\alpha}. \end{aligned}$$

Choose the various constants in the definition (4.19), (4.20), and (4.21) of Υ in the following order: K, $K^\flat$, $K^\sharp$, and $\mathcal{K}$ sufficiently large; ε_o and δ sufficiently small. A possible choice is

$$K = 1 + 24C, \qquad K^\flat = K^\sharp = 1 + 8CK, \qquad \mathcal{K} = 1 + 2C(1 + K + KK^\flat), \qquad \varepsilon_o < \frac{1}{3}$$

with $\delta < \frac{4}{(KK^\flat)^2}$.

(1) Interaction between small waves of different families.

Assume that the interaction takes place in the leftmost phase. Compute first

$$\begin{aligned} \Delta V^\flat_\xi &= \sum_{i,\alpha} p^{\flat +}_{i,\alpha} \left|\sigma^+_{i,\alpha}\xi^+_{i,\alpha}\right| - \sum_i p^{\flat -}_i \left|\sigma^-_i \xi^-_i\right| \\ &\leq CK(1+\varepsilon_o)\left|\sigma_1^-\sigma_2^-\right| \left(\left|\xi_1^-\right| + \left|\xi_2^-\right|\right) \end{aligned}$$

since $p^{\flat +}_{i,\alpha} = p^{\flat -}_i$ for every α and i. Moreover,

$$\begin{aligned} \Delta Q^\flat_\xi = &\sum_{i,\alpha} \sum_{(j,\beta)\in\mathcal{A}^{\flat +}_{i,\alpha}} \left|\sigma^+_{i,\alpha}\sigma_{j,\beta}\right| \left(p^{\flat +}_{i,\alpha}\left|\xi^+_{i,\alpha}\right| + p^\flat_{j,\beta}\left|\xi_\beta\right|\right) \\ &- \sum_i \sum_{(j,\beta)\in\mathcal{A}^{\flat -}_i} \left|\sigma^-_i\sigma_{j,\beta}\right| \left(p^{\flat -}_i\left|\xi^-_i\right| + p^\flat_{j,\beta}\left|\xi_\beta\right|\right) - \left|\sigma_1^-\sigma_2^-\right| \sum_i p^{\flat -}_i \left|\xi^-_i\right| \\ \leq &\sum_i p^{\flat -}_i \left|\sum_\alpha \left|\sigma^+_{i,\alpha}\xi^+_{i,\alpha}\right| - \left|\sigma^-_i\xi^-_i\right|\right| V^{\flat -} \\ &+ \sum_i p^{\flat -}_i \left|\sum_\alpha \left|\sigma^+_{i,\alpha}\right| - \left|\sigma^-_i\right|\right| V^{\flat -}_\xi - \frac{2}{3}\left|\sigma_1^-\sigma_2^-\right| \left(\left|\xi_1^-\right| + \left|\xi_2^-\right|\right) \\ \leq &\, CK(1+\varepsilon_o)\left|\sigma_1^-\sigma_2^-\right| \left(\left|\xi_1^-\right| + \left|\xi_2^-\right|\right)\delta \\ &+ C\left|\sigma_1^-\sigma_2^-\right| \left(\left|\sigma_1^-\right| + \left|\sigma_2^-\right|\right) V^{\flat -}_\xi - \frac{2}{3}\left|\sigma_1^-\sigma_2^-\right| \left(\left|\xi_1^-\right| + \left|\xi_2^-\right|\right) \\ \leq &\, CK\left|\sigma_1^-\sigma_2^-\right| \left(\left|\sigma_1^-\right| + \left|\sigma_2^-\right|\right) V^{\flat -}_\xi - \frac{1}{2}\left|\sigma_1^-\sigma_2^-\right| \left(\left|\xi_1^-\right| + \left|\xi_2^-\right|\right). \end{aligned}$$

At last

$$\Delta Q^\flat \leq -\frac{1}{2}\left|\sigma_1^- \sigma_2^-\right| \ .$$

The above estimates are valid under the choices above of the weights and for δ sufficiently small. From the previous estimates, it follows that

$$\begin{aligned}
\Delta \Upsilon_\xi^\flat &= \Delta V_\xi^\flat + K^\flat \Delta Q_\xi^\flat + Q^{\flat+} \Delta V_\xi^\flat + V_\xi^{\flat-} \Delta Q^\flat \\
&\leq \left(CK(1+\varepsilon_o)(1+\delta) - \frac{1}{2}K^\flat \right) \left|\sigma_1^- \sigma_2^-\right| \left(\left|\xi_1^-\right| + \left|\xi_2^-\right| \right) \\
&\qquad + \left(CK^\flat K\delta - \frac{1}{2} \right) \left|\sigma_1^- \sigma_2^-\right| V_\xi^{\flat-} \\
&\leq 0 \ .
\end{aligned} \tag{5.19}$$

By Proposition 4.1 we obtain

$$\Delta \Upsilon_\xi \leq \Delta \Upsilon_\xi^\flat e^{K\Upsilon^-} \leq 0.$$

Analogous estimates hold for interactions of waves of different families in the rightmost phase.

(2) Interaction between two shocks of the same family.

Again, assume that the interaction takes place on the leftmost phase. Similar to the previous case we find

$$\Delta V_\xi^\flat \leq CK(1+\varepsilon_o)|\sigma'\sigma''| \left(|\xi'| + |\xi''|\right),$$

while

$$\begin{aligned}
\Delta Q_\xi^\flat = &\sum_{(j,\beta)\in\mathcal{A}_1^{\flat+}} \left|\sigma_1^+ \sigma_{j,\beta}\right| \left(p_1^{\flat+} \left|\xi_1^+\right| + p_{j,\beta}^\flat |\xi_\beta| \right) \\
&+ \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{2,\alpha}^{\flat+}} \left|\sigma_{2,\alpha}^+ \sigma_{j,\beta}\right| \left(p_{2,\alpha}^{\flat+} \left|\xi_{2,\alpha}^+\right| + p_{j,\beta}^\flat |\xi_\beta| \right) \\
&- \sum_{(j,\beta)\in\mathcal{A}'} |\sigma' \sigma_{j,\beta}| \left(p'|\xi'| + p_{j,\beta}^\flat |\xi_\beta| \right) - \sum_{(j,\beta)\in\mathcal{A}''} |\sigma'' \sigma_{j,\beta}| \left(p''|\xi''| + p_{j,\beta}^\flat |\xi_\beta| \right) \\
&-|\sigma'\sigma''| \left(p'|\xi'| + p''|\xi''| \right) \\
\leq &\, K \left((1-\varepsilon_o) \left| \left|\sigma_1^+ \xi_1^+\right| - |\sigma'\xi'| - |\sigma''\xi''| \right| + (1+\varepsilon_o) \sum_\alpha \left|\sigma_{2,\alpha}^+ \xi_{2,\alpha}^+\right| \right) V^{\flat-} \\
&+K \left((1-\varepsilon_o) \left|\sigma_1^+ - (\sigma' + \sigma'')\right| + (1+\varepsilon_o) \sum_\alpha \left|\sigma_{2,\alpha}^+\right| \right) V_\xi^{\flat-} \\
&-\frac{2}{3}|\sigma'\sigma''| \left(|\xi'| + |\xi''|\right) \\
\leq &\, CK(1+\varepsilon_o)|\sigma'\sigma''| \left(|\xi'| + |\xi''|\right) V^{\flat-} + CK|\sigma'\sigma''| \left(|\sigma'| + |\sigma''|\right) V_\xi^{\flat-} \\
&-\frac{2}{3}|\sigma'\sigma''| \left(|\xi'| + |\xi''|\right) \\
\leq &\, CK|\sigma'\sigma''| \left(|\sigma'| + |\sigma''|\right) V_\xi^{\flat-} - \frac{1}{2}|\sigma'\sigma''| \left(|\xi'| + |\xi''|\right).
\end{aligned}$$

Since

$$\Delta Q^\flat \leq -\frac{1}{2}|\sigma'\sigma''|,$$

in this case the proof is completed exactly as in Step 1.

(3) Interaction between two small waves of the same family but of different sign.

Assume that $\sigma' > 0$, $\sigma'' < 0$ and that both belong to the first family, the other situations being similar. Observe that in this case one must have $\sigma_1^+ \leq 0$, otherwise the two incoming wave-fronts would have exactly the same speed. We have

$$\begin{aligned}\Delta V_\xi^\flat &\leq K(1-\varepsilon_o)\left|\sigma_1^+\xi_1^+\right| + \sum_\alpha p_{2,\alpha}^+\left|\sigma_{2,\alpha}^+\xi_{2,\alpha}^+\right| - (1+\varepsilon_o)|\sigma'\xi'| - (1-\varepsilon_o)|\sigma''\xi''|\\ &\leq CK(1+\varepsilon_o)|\sigma'\sigma''|\left(|\xi'|+|\xi''|\right) - 2\varepsilon_o|\sigma'\xi'|\ .\end{aligned}$$

For the interaction potential, we find

$$\begin{aligned}\Delta Q_\xi^\flat = & \sum_{(j,\beta)\in\mathcal{A}_1^{\flat+}} \left(\left|\sigma_1^+\sigma_{j,\beta}\right|\left(p_1^{\flat+}\left|\xi_1^+\right| + p_{j,\beta}^\flat|\xi_\beta|\right)\right)\\ & + \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{2,\alpha}^{\flat+}} \left(\left|\sigma_{2,\alpha}^+\sigma_{j,\beta}\right|\left(p_{2,\alpha}^{\flat+}\left|\xi_{2,\alpha}^+\right| + p_{j,\beta}^\flat|\xi_\beta|\right)\right)\\ & - \sum_{(j,\beta)\in\mathcal{A}'} \left(\left|\sigma'\sigma_{j,\beta}\right|\left(p'|\xi'| + p_{j,\beta}^\flat|\xi_\beta|\right)\right)\\ & - \sum_{(j,\beta)\in\mathcal{A}''} \left(\left|\sigma''\sigma_{j,\beta}\right|\left(p''|\xi''| + p_{j,\beta}^\flat|\xi_\beta|\right)\right) - |\sigma'\sigma''|\left(p'|\xi'|+p''|\xi''|\right)\\ \leq & K\left((1-\varepsilon_o)\left|\left|\sigma_1^+\xi_1^+\right| - |\sigma''\xi''|\right| + (1+\varepsilon_o)\sum_\alpha\left|\sigma_{2,\alpha}^+\xi_{2,\alpha}^+\right|\right)V^{\flat-}\\ & + K(1+\varepsilon_o)\sum_\alpha\left|\sigma_{2,\alpha}^+\right|V_\xi^{\flat-} - \frac{2}{3}|\sigma'\sigma''|\left(|\xi'|+|\xi''|\right)\\ \leq & CK|\sigma'\sigma''|\left(|\xi'|+|\xi''|\right)V^{\flat-} + K|\sigma'\xi'|V^{\flat-} + CK|\sigma'\sigma''|\left(|\sigma'|+|\sigma''|\right)V_\xi^{\flat-}\\ & - \frac{2}{3}|\sigma'\sigma''|\left(|\xi'|+|\xi''|\right)\\ \leq & CK|\sigma'\sigma''|\left(|\sigma'|+|\sigma''|\right)V_\xi^{\flat-} + K|\sigma'\xi'|V^{\flat-} - \frac{1}{2}|\sigma'\sigma''|\left(|\xi'|+|\xi''|\right).\end{aligned}$$

Since the inequality $\Delta Q^\flat \leq -|\sigma'\sigma''|/2$ still holds, using (5.19), we have

$$\begin{aligned}\Delta\Upsilon_\xi^\flat \leq & \left(CK(1+\varepsilon_o)(1+\delta) - \frac{1}{2}K^\flat\right)|\sigma'\sigma''|\left(|\xi'|+|\xi''|\right)\\ & + \left(CKK^\flat\delta - \frac{1}{2}\right)|\sigma'\sigma''|V_\xi^{\flat-} + \left(KK^\flat\delta - 2\varepsilon_o\right)|\sigma'\xi'|\\ \leq & \ 0,\end{aligned}\tag{5.20}$$

so that $\Delta\Upsilon_\xi \leq 0$ for δ sufficiently small.

(4) Interaction between the phase boundary and a small wave.

We consider the case of a 2-wave hitting the phase boundary from the left, the other case being similar (see Figure 5.2(ii)). By (5.15) and (5.3) we deduce

$$\Delta V_\xi^\flat = \sum_\alpha p_{1,\alpha}^{\flat+}\left|\sigma_{1,\alpha}^+\xi_{1,\alpha}^+\right| - p_2^{\flat-}\left|\sigma_2^-\xi_2^-\right|$$

$$
\begin{aligned}
&\le \left(C(1+\varepsilon_o) - K(1-\varepsilon_o)\right) |\sigma_2^- \xi_2^-| + 2C|\sigma_2^-||\hat{\xi}^-| \\
&\le \left(2C - \frac{1}{2}K\right) |\sigma_2^- \xi_2^-| + 2C|\sigma_2^-|\left|\hat{\xi}^-\right|,
\end{aligned}
$$

where we used the above choice of K.

$$
\begin{aligned}
\Delta Q_\xi^\flat &= \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{1,\alpha}^{\flat+}} |\sigma_{1,\alpha}^+ \sigma_{j,\beta}| \left(p_{1,\alpha}^{\flat+}|\xi_{1,\alpha}^+| + p_{j,\beta}^\flat |\xi_{j,\beta}|\right) \\
&\le \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{1,\alpha}^{\flat+}} p_{1,\alpha}^{\flat+}|\sigma_{1,\alpha}^+ \sigma_{j,\beta}||\xi_{1,\alpha}^+| + \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{1,\alpha}^{\flat+}} p_{j,\beta}^\flat |\sigma_{1,\alpha}^+ \sigma_{j,\beta}||\xi_{j,\beta}| \\
&\le 2C|\sigma_2^-|\left(|\xi_2^-| + |\hat{\xi}^-|\right) V^{\flat-} + 2CK|\sigma_2^-|V_\xi^{\flat-}\ , \\
\Delta Q^\flat &= \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{1,\alpha}^{\flat+}} |\sigma_{1,\alpha}^+ \sigma_{j,\beta}| \\
&\le C|\sigma_2^-|V^{\flat-}\ .
\end{aligned}
$$

By (5.19) it follows that

$$
\begin{aligned}
\Delta \Upsilon_\xi^\flat \le &\left(2C(1+2\delta) - \frac{1}{2}K\right) |\sigma_2^- \xi_2^-| + 2C(1+2\delta)|\sigma_2^-|\left|\hat{\xi}^-\right| \\
&+2C(1+K)\left|\sigma_2^-\right|V_\xi^{\flat-}\ . \qquad (5.21)
\end{aligned}
$$

We now consider the terms referring to the right of the phase boundary. By (5.16) and (5.3) we have

$$
\begin{aligned}
\Delta V_\xi^\sharp &= \sum_\alpha p_{2,\alpha}^{\sharp+} |\sigma_{2,\alpha}^+ \xi_{2,\alpha}^+| \\
&\le 2C|\sigma_2^-|\left(|\xi_2^-| + |\hat{\xi}^-|\right), \\
\Delta Q_\xi^\sharp &= \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{2,\alpha}^{\sharp+}} |\sigma_{2,\alpha}^+ \sigma_{j,\beta}| \left(p_{2,\alpha}^{\sharp+}|\xi_{2,\alpha}^+| + p_{j,\beta}^\sharp |\xi_{j,\beta}|\right) \\
&\le 2CK|\sigma_2^-|\left(|\xi_2^-| + |\hat{\xi}^-|\right) V^{\sharp-} + CK|\sigma_2^-|V_\xi^{\sharp-}\ , \\
\Delta Q^\sharp &= \sum_\alpha \sum_{(j,\beta)\in\mathcal{A}_{2,\alpha}^{\sharp+}} |\sigma_{2,\alpha}^+ \sigma_{j,\beta}| \\
&\le C|\sigma_2^-|V^{\sharp-}\ .
\end{aligned}
$$

Therefore,

$$
\Delta \Upsilon_\xi^\sharp \le 2C\left(1 + (1+KK^\sharp)\delta\right)|\sigma_2^-|\left(|\xi_2^-| + |\hat{\xi}^-|\right) + C(\delta + KK^\sharp)|\sigma_2^-|V_\xi^{\sharp-}\ . \qquad (5.22)
$$

From the inequalities (5.21), (5.22), (5.14), and (5.11), it follows finally that

$$
\begin{aligned}
\Delta \Upsilon_\xi &= \left(\Delta \Upsilon_\xi^\flat + \Delta \Upsilon_\xi^\sharp + \Delta|\hat{\xi}|\right) e^{\mathcal{K}\Upsilon^+} + \left(\Upsilon_\xi^{\flat-} + \Upsilon_\xi^{\sharp-} + |\hat{\xi}^-|\right)\left(e^{\mathcal{K}\Upsilon^+} - e^{\mathcal{K}\Upsilon^-}\right) \\
&\le \left(\Delta \Upsilon_\xi^\flat + \Delta \Upsilon_\xi^\sharp + \Delta|\hat{\xi}| - \mathcal{K}|\Delta\Upsilon|\left(\Upsilon_\xi^{\flat-} + \Upsilon_\xi^{\sharp-} + |\hat{\xi}^-|\right)\right) e^{\mathcal{K}\Upsilon^+} \\
&\le \Big[\left(2C(3 + 3\delta + KK^\sharp\delta) - \frac{1}{2}K\right)|\sigma_2^- \xi_2^-|
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \left(2C(3+3\delta+KK^\sharp\delta)-\mathcal{K}\right)\left|\sigma_2^-\right|\left|\hat{\xi}^-\right| \\
&\quad + \left(2C(1+K)-\mathcal{K}\right)\left|\sigma_2^-\right|V_\xi^{\flat -} + \left(C(\delta+KK^\sharp)-\mathcal{K}\right)\left|\sigma_2^-\right|V_\xi^{\sharp -}\Big]e^{\mathcal{K}\Upsilon^+} \\
&\leq 0
\end{aligned}
$$

by the above choices of the weights. The proposition is proved in Case 1.

Case 2. Let us choose the weights in (4.22) and (4.24). Define

$$
\begin{aligned}
p_{1,\alpha}^\flat &= 1+\varepsilon_o \operatorname{sign}\sigma_{1,\alpha}, \qquad & p_{2,\alpha}^\flat &= K_2^\flat(1+\varepsilon_o \operatorname{sign}\sigma_{2,\alpha}), \\
p_{i,\alpha}^\natural &= K_i^\natural(1+\varepsilon_o \operatorname{sign}\sigma_{1,\alpha}), & & \\
p_{1,\alpha}^\sharp &= K_1^\sharp(1+\varepsilon_o \operatorname{sign}\sigma_{1,\alpha}), & p_{2,\alpha}^\sharp &= 1+\varepsilon_o \operatorname{sign}\sigma_{2,\alpha}.
\end{aligned}
$$

Choose first $K_1^\natural$ and $K_2^\natural$ such that

$$((1+\varepsilon_o)(1+C\delta))\left|R_\natural^\flat\Theta^\flat\right|\cdot K_2^\natural-(1-\varepsilon_o)K_1^\natural < 3(1+\varepsilon_o)C, \tag{5.23}$$

$$((1+\varepsilon_o)(1+C\delta))\left|R_\natural^\sharp\Theta^\sharp\right|\cdot K_1^\natural-(1-\varepsilon_o)K_2^\natural < 3(1+\varepsilon_o)C, \tag{5.24}$$

which is possible provided $\varepsilon_o = \sqrt{\delta}$ and δ sufficiently small, thanks to the strong nonresonance condition (2.13). The other weights are chosen similarly to the previous case, i.e., in the order $K_2^\flat = K_1^\sharp$, $K^\flat$, $K^\natural$, $K^\sharp$, and $\mathcal{K}$. As usual, choose finally a suitably small δ.

The estimates in this case differ from the previous ones only in the interaction of a wave coming from the middle phase against a phase boundary. We now consider only this interaction in detail.

$$
\begin{aligned}
\Delta\Upsilon_\xi^\flat &= C(1+\varepsilon_o)\left(1+\delta+K^\flat\delta\right)\left|\sigma_1^-\right|\left(\left|\xi_1^-\right|+\left|\xi^{\flat -}\right|\right)+C(\delta+K^\flat)\left|\sigma_1^-\right|V_\xi^{\flat -}, \\
\Delta\Upsilon_\xi^\natural &= \left((1+\varepsilon_o)(1+C\delta)\left|R_\natural^\flat\Theta^\flat\right|K_2^\natural-(1-\varepsilon_o)K_1^\natural\right)\left|\sigma_1^-\xi_1^-\right| \\
&\qquad +C(1+\varepsilon_o)K_2^\natural\left|\sigma_1^-\right|\left|\xi^{\flat -}\right|+(1+C\delta)R_\natural^\flat K_2^\natural\left|\sigma_1^-\right|V_\xi^{\natural -}, \\
\Delta\Upsilon_\xi &\leq \left(\Delta\Upsilon_\xi^\flat+\Delta\Upsilon_\xi^\natural+\Delta\left|\xi^\flat\right|-\mathcal{K}\left|\sigma_1^-\right|V_\xi^{\flat -}-\mathcal{K}\left|\sigma_1^-\right|V_\xi^{\natural -}-\mathcal{K}\left|\sigma_1^-\right|\left|\xi^{\flat -}\right|\right)e^{\mathcal{K}\Upsilon^+} \\
&\leq \left(C(1+\varepsilon_o)\left(1+\delta+K^\flat\delta\right)+C+(1+\varepsilon_o)(1+C\delta)\left|R_\natural^\flat\Theta^\flat\right|K_2^\natural-(1-\varepsilon_o)K_1^\natural\right) \\
&\qquad \times\left|\sigma_1^-\xi_1^-\right| \\
&\qquad +\left(C(1+\varepsilon_o)\left(1+\delta+K^\flat\delta\right)+C(1+\varepsilon_o)K_2^\natural+C-\mathcal{K}\right)\left|\sigma_1^-\right|\left|\xi^{\flat -}\right| \\
&\qquad +\left(C(\delta+K^\flat)-\mathcal{K}\right)\left|\sigma_1^-\right|V_\xi^{\flat -}+\left(1+C\delta R_\natural^\flat K_2^\natural-\mathcal{K}\right)\left|\sigma_1^-\right|V_\xi^{\natural -} \\
&\leq 0.
\end{aligned}
$$

This concludes the proof of Proposition 4.7. □

Proof of Proposition 4.8. Let $u', u'' \in \mathcal{D}_{\delta'}^\varepsilon$ be given, with $\delta' > 0$ small. We consider only case (2), since case (1) is simpler.

For $i = 1, 2$, call x_i', x_i'' the positions of the phase boundaries in u' and u'', respectively. One can then connect u' with u'' in such a way that each intermediate state u^θ contains exactly two large phase boundaries. For example, assume $x_1' < x_1'' <$

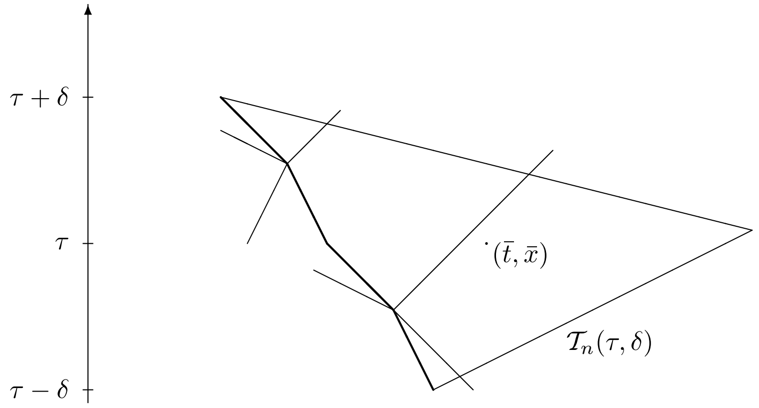


FIG. 5.3. *The set $\mathcal{T}_n(\tau,\delta)$.*

$x'_2 < x''_2$, the other cases being entirely similar. We first define the path $\gamma_1\colon [x'_2, x''_2] \mapsto \mathcal{D}^{\varepsilon}_{\delta}$,

$$\gamma_1(\theta) = u' \cdot \chi_{]-\infty,x'_2]\cup]\theta\ \ ,\infty+[} + u'' \cdot \chi_{]x'_2,\theta]}$$

joining u' with the intermediate function

$$w(x) = \begin{cases} u'(x) & \text{if } x \in\ ]-\infty, x'_2] \cup]x''_2, \infty[, \\ u''(x) & \text{if } x \in\ ]x'_2, x''_2]. \end{cases}$$

We then connect w with u'' by setting

$$\gamma_2(\theta) = u'' \cdot \chi_{]-\infty,\theta]} + w \cdot \chi_{]\theta,+\infty[}\,.$$

The concatenation of γ_1 and γ_2 yields the desired path. □

Proof of Proposition 4.10. As we announced in section 4 we give only the proof of the Ψ-admissibility of the solutions.

Fix a positive sequence $\{\varepsilon_n\colon n \in \mathbf{N}\}$ converging to 0. For all n, let u_n be the ε_n-approximate solution constructed by the algorithm. Call $x = \Lambda_n(t)$ the equation of the leftmost approximate phase boundary in u_n; the other cases are completely analogous.

Then choose a positive τ. For all n and positive δ, define $\mathcal{T}_n(\tau,\delta)$ as the following region of the (t,x)-plane (see Figure 5.3):

$$\begin{aligned} &\mathcal{T}_n(\tau,\delta) \\ &= \left\{ (t,x) \in [0,+\infty[\, \times \mathbf{R}\colon \left\{ \begin{array}{l} t \in]\tau-\delta, \tau+\delta[, \\ x < \Lambda_n(\tau-\delta) + \lambda^{\min} \cdot (t-(\tau-\delta)), \\ x < \Lambda_n(\tau+\delta) - \lambda^{\min} \cdot (t-(\tau+\delta)), \\ x > \Lambda_n(t) \end{array} \right. \right\}. \end{aligned} \tag{5.25}$$

Due to the subsonic hypothesis (2.4), $\mathcal{T}_n(\tau,\delta)$ is bounded. Moreover, note that by (2.1), 1-waves may exit $\mathcal{T}_n(\tau,\delta)$ only by crossing the phase boundary, while they may enter $\mathcal{T}_n(\tau,\delta)$ only through the bottom right side. Similarly, 2-waves may enter $\mathcal{T}_n(\tau,\delta)$ only by crossing the phase boundary, while they may exit $\mathcal{T}_n(\tau,\delta)$ only crossing the top right side. Due to (iii) in Proposition 4.1 there is only a finite number (depending on n) of such exiting or entering waves. Clearly, we can have creation or cancellation of both 1- and 2-waves inside $\mathcal{T}_n(\tau,\delta)$. Let us call $\mathcal{T}(\tau,\delta)$ the region analogous to (5.25) but constructed with reference to the exact phase boundary $x = \Lambda(t)$.

The following lemma is crucial; it is analogous to the celebrated Lemma 3.4 of [10] but we emphasize that its proof is much easier, as a consequence of the wave-front tracking scheme.

LEMMA 5.1. *For all but countably many τ and eventually passing to a subsequence of the approximate solutions, the following holds: for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon,\tau)$ such that for every $(t,x) \in \mathcal{T}(\tau,\delta)$ and for all n large*

$$|u_n(t,x) - u_n\left(\tau,\Lambda_n(\tau)+\right)| < \varepsilon\ .$$

Proof. Let $\Upsilon_n(t) = \Upsilon^{\varepsilon_n}\left(u_n(t,\cdot)\right)$. By Helly's theorem, eventually passing to a subsequence we define for all $t \geq 0$

$$\tilde{\Upsilon}(t) = \lim_{n\to+\infty} \Upsilon_n(t).$$

Fix $\varepsilon > 0$ so small that for all the future constants C the estimate $C\varepsilon < 1$ holds. For every τ apart from a finite set of times (dependent on ε), there exists a positive δ such that

$$\left|\tilde{\Upsilon}(\tau+\delta+) - \tilde{\Upsilon}(\tau-\delta-)\right| < \varepsilon^6\ .$$

In Case 2 we eventually further restrict δ to ensure that the other phase boundary does not intersect $\mathcal{T}(\tau,\delta)$.

Then there exists a positive sequence $\{\delta_n : n \in \mathbf{N}\}$ such that δ_n increases, $\delta_n \to \delta$, and

$$|\Upsilon_n(\tau+\delta_n+) - \Upsilon_n(\tau-\delta_n-)| < \varepsilon^5 \tag{5.26}$$

for n sufficiently large. Choose an arbitrary $(\bar{t},\bar{x}) \in \mathcal{T}(\tau,\delta)$ and consider n sufficiently large in order that $(\bar{t},\bar{x}) \in \mathcal{T}_n(\tau,\delta_n)$ and (5.26) holds.

To compute $|u_n(\bar{t},\bar{x}) - u_n\left(\tau,\Lambda_n(\tau)+\right)|$, first draw the horizontal line segment $\mathcal{S}_n$ joining $(\bar{t},\bar{x})$ with the approximate phase boundary $x = \Lambda_n(t)$. Then, using the triangle inequality and the known estimate on $\Delta\Upsilon_n$

$$\begin{aligned}
&|u_n(\bar{t},\bar{x}) - u_n\left(\tau,\Lambda_n(\tau)+\right)| \\
&\leq |u_n(\bar{t},\bar{x}) - u_n\left(\bar{t},\Lambda_n(\bar{t})+\right)| + |u_n\left(\bar{t},\Lambda_n(\bar{t})+\right) - u_n\left(\tau,\Lambda_n(\tau)+\right)| \\
&\leq \mathrm{TV}\left(u_n\left(\bar{t},x\right) : x \in [\Lambda_n(\bar{t}),\bar{x}]\right) \\
&\quad + \mathrm{TV}\left(u_n\left(t,\Lambda_n(t)+\right) : t \in [\tau-\delta_n,\tau+\delta_n]\right) \\
&\leq C \sum_{\alpha:\,\sigma_{i,\alpha}\text{crosses } \mathcal{S}_n} |\sigma_{i,\alpha}| + C|\Upsilon_n(\tau+\delta_n+) - \Upsilon_n(\tau-\delta_n-)| \\
&\leq C \sum_{\alpha:\,\sigma_{i,\alpha}\text{crosses } \mathcal{S}_n} |\sigma_{i,\alpha}| + \varepsilon^4.
\end{aligned}$$

The rest of the proof aims at bounding from above the total quantity of waves crossing $\mathcal{S}_n$ and will be achieved by the following two claims.

Claim 1. For all large n, the total size of 1-waves exiting $\mathcal{T}_n(\tau,\delta_n)$ and the total size of 2–waves entering $\mathcal{T}_n(\tau,\delta_n)$ are both lower than ε^4.

Assume that $\sigma_{1,\alpha}^+$ exits $\mathcal{T}_n(\tau,\delta_n)$ at some time $t_\alpha \in [\tau-\delta_n, \tau+\delta_n]$. By (5.3) and (5.7) or (5.11) in Case 2

$$\sum_\alpha |\sigma_{1,\alpha}^+| \le C \sum_\alpha \left(|\sigma_{1,\alpha}^-| + |\sigma_{2,\alpha}^-|\right) \le C \sum_\alpha |\Delta \Upsilon_n(t_\alpha)| \le C\varepsilon^5 \le \varepsilon^4$$

by (5.26) and the decreasing of Υ_n. Here $\sigma_{1,\alpha}^-$, $\sigma_{2,\alpha}^-$ are the waves hitting the phase boundary.

The case of 2-waves is entirely analogous.

Claim 2. For all large n, the total size of 1- and 2-waves crossing $\mathcal{S}_n$ is smaller than ε^3.

As above, let $\sigma_{1,\alpha}$ for $\alpha = 1,\ldots,$ be the size of the 1-waves crossing $\mathcal{S}_n$. By eventually prolonging $\sigma_{1,\alpha}$ as a null wave, we can assume that $\sigma_{1,\alpha}$ exits $\mathcal{T}_n(\tau,\delta_n)$ through the phase boundary at some time $t_\alpha \in [\tau-\delta_n,\tau+\delta_n]$. By (4.10) in [5] (which generalizes to the case of many colliding waves our estimates (5.1) and (5.2)), by the definitions (4.8)–(4.10) or (4.13)–(4.14), by the interaction estimates (5.5) and (5.6), and by the previous Claim 1

$$\begin{aligned}\sum_{\alpha\colon \sigma_{1,\alpha} \text{crosses } \mathcal{S}_n} |\sigma_{1,\alpha}| &\le \sum_\alpha |\sigma_{1,\alpha}(t_\alpha+)| + \sum_{\substack{t_h \text{ interaction time}\\ \bar t \le t_h < t_\alpha}} |\Delta Q(t_h)| \\ &\le \varepsilon^4 + C \sum_{\substack{t_h \text{ interaction time}\\ \bar t \le t_h < t_\alpha}} |\Delta \Upsilon_n(t_h)| \\ &\le \varepsilon^4 + C\cdot |\Upsilon_n(\tau+\delta+) - \Upsilon_n(\tau-\delta-)| \\ &\le \varepsilon^3\end{aligned}$$

for a suitable constant C, where Q stands for $Q^\sharp$ in Case 1 and for $Q^\natural$ in Case 2.

For 2-waves the proof is entirely symmetric, the only difference being that now waves are prolonged as null waves backwards.

The lemma is therefore proved. □

The admissibility of the solutions follows from this lemma, since we have, arguing as in [10],

$$\lim_{n\to+\infty} u_n\big(t,\Lambda_n(t)+\big) = u\big(t,\Lambda(t)+\big).$$

This concludes the proof of Proposition 4.10. □

Acknowledgment. The authors thank the Faculty of Sciences in Varese for the hospitality provided to both the authors.

REFERENCES

[1] R. Abeyaratne and J. K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Rational Mech. Anal., 114 (1991), pp. 119–154.

[2] A. Bressan, *The unique limit of the Glimm scheme*, Arch. Rational Mech. Anal., 130 (1995), pp. 205–230.

[3] A. Bressan, *The semigroup approach to systems of conservation laws*, Mat. Contemp., 10 (1996), pp. 21–74.

[4] A. Bressan and R. M. Colombo, *The semigroup generated by* 2×2 *conservation laws*, Arch. Rational Mech. Anal., 133 (1995), pp. 1–75.

[5] A. Bressan and R. M. Colombo, *Unique solutions of* 2×2 *conservation laws with large data*, Indiana Univ. Math. J., 44 (1995), pp. 677–725.

[6] A. Corli, *Noncharacteristic phase boundaries for general systems of conservation laws*, Ital. J. Pure Appl. Math., to appear.

[7] A. Corli and M. Sablé-Tougeron, *Kinetic stabilization of a nonlinear sonic phase boundary*, Arch. Rational Mech. Anal., to appear.

[8] H. Fan, *A limiting viscosity approach to the Riemann problem for materials exhibiting a change of phase* (II), Arch. Rational Mech. Anal., 116 (1992), pp. 317–337.

[9] H. Fan, *Global versus local admissibility criteria for dynamic phase boundaries*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 927–944.

[10] J. Glimm and P. D. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc., 101 (1970).

[11] R. D. James, *The propagation of phase boundaries in elastic bars*, Arch. Rational Mech. Anal., 73 (1980), pp. 125–158.

[12] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[13] Ph. LeFloch, *Propagating phase boundaries: Formulation of the problem and existence via the Glimm method*, Arch. Rational Mech. Anal., 123 (1993), 153–197.

[14] Y. Lin, *A Riemann problem for an elastic bar that changes phase*, Quart. Appl. Math., 53 (1995), pp. 575–600.

[15] T.-P. Liu, *The Riemann problem for general* 2×2 *conservation laws*, Trans. Amer. Math. Soc., 199 (1974), pp. 89–112.

[16] T.-P. Liu, *Nonlinear stability and instability of overcompressive shock waves*, in Shock Induced Transitions and Phase Structures in General Media, J. E. Dunn, R. Fosdick, and M. Slemrod, eds., Springer, New York, 1993, pp. 159–167.

[17] M. Sablé-Tougeron, *Méthode de Glimm et problème mixte*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 423–443.

[18] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[19] M. Slemrod, *A limiting viscosity approach to the Riemann problem for materials exhibiting change of phase*, Arch. Rational Mech. Anal., 105 (1989), pp. 327–365.

[20] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.

[21] L. Truskinovsky, *Kinks versus shocks*, in Shock Induced Transitions and Phase Structures in General Media, J. E. Dunn, R. Fosdick, and M. Slemrod, eds., Springer, New York, 1993, pp. 185–229.

[22] L. Truskinovsky, *About the "normal growth" approximation in the dynamical theory of phase transitions*, Contin. Mech. Thermodyn., 6 (1994), pp. 185–208.

[23] I. M. Ward, *Mechanical Properties of Solid Polymers*, Wiley, New York, 1971.

SIAM J. Math. Anal.
Vol. 31, No. 1, pp. 63–79

ON THE ROLE OF MEAN CURVATURE IN SOME SINGULARLY PERTURBED NEUMANN PROBLEMS[*]

MANUEL DEL PINO[†], PATRICIO L. FELMER[†], AND JUNCHENG WEI[‡]

Abstract. We construct solutions exhibiting a single spike-layer shape around some point of the boundary as $\varepsilon \to 0$ for the problem

$$\begin{cases} \varepsilon^2 \triangle u - u + u^p = 0 \quad \text{in } \Omega, \\ u > 0 \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{0.1}$$

where Ω is a bounded domain with smooth boundary in R^N, $p > 1$, and $p < \frac{N+2}{N-2}$ if $N \geq 3$. Our main result states that given a *topologically nontrivial critical point* of the mean curvature function of $\partial\Omega$, for instance, a possibly degenerate local maximum, local minimum, or saddle point, there is a solution with a single local maximum, which is located at the boundary and approaches this point as $\varepsilon \to 0$ while vanishing asymptotically elsewhere.

Key words. spike layer, singular perturbations, Neumann problems

AMS subject classifications. 35B25, 35J20, 35B40

PII. S0036141098332834

1. Introduction.

In this paper, we are concerned with the following singularly perturbed problem:

$$\begin{cases} \varepsilon^2 \Delta u - u + u^p = 0 \quad \text{in } \Omega, \\ u > 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial v} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset R^N$ is a smooth, not necessarily bounded domain; $\varepsilon > 0$; and $1 < p < (N+2)/(N-2)$ if $N \geq 3$ and $p > 2$ if $N = 2$.

Equation (1.1) arises from various applications. For instance, it can be regarded as that satisfied by stationary solutions for the Keller–Segal system in chemotaxis (see [14], [17], [19]) and the Gierer–Meinhardt system in biological pattern formation (see [12], [21]).

In [17], Lin, Ni, and Takagi first studied the problem of existence of least-energy solutions. Subsequently, Ni and Takagi in [19] and [21] showed that the least-energy solution u_ε has a unique local maximum point P_ε, which is located on $\partial\Omega$. Moreover, $u_\varepsilon \to 0$ in $C^1_{loc}(\overline{\Omega}\backslash P_\varepsilon)$ and $u_\varepsilon(P_\varepsilon) \to \alpha > 0$ as $\varepsilon \to 0$. Such a family of solutions is usually called a *boundary spike-layer*. Moreover, they are able to locate the spike by establishing that P_ε approaches the *most curved* part of $\partial\Omega$, namely, $H(P_\varepsilon) \to$

[*] Received by the editors January 22, 1998; accepted for publication (in revised form) March 5, 1999; published electronically November 10, 1999.

http://www.siam.org/journals/sima/31-1/33283.html

[†] Departamento de Ingeniería Matemática F.C.F.M., Universidad de Chile, Casilla 170 Correo 3, Santiago, Chile (delpino@dim.uchile.cl, pfelmer@dim.uchile.cl). The first and second authors were supported by Cátedra Presidencial, FONDAP de Matemáticas Aplicadas, and FONDECYT grants 1960698 and 1970775.

[‡] Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). This author was supported by an Earmarked grant of RGC of Hong Kong.

$\max_{P\in\partial\Omega} H(P)$, where H is the mean curvature. Later Wei studied general boundary spike solutions in [23] and showed that for any solution with single peak P_ε on $\partial\Omega$, $\nabla_{\tau_{P_\varepsilon}} H(P_\varepsilon) \to 0$, where $\nabla_{\tau_{P_\varepsilon}}$ denote the tangential gradient at $P_\varepsilon \in \partial\Omega$. On the other hand, if $P_0 \in \partial\Omega$, $\nabla_{\tau_{P_0}} H(P_0) = 0$ and the matrix $(\nabla^2_{\tau_{P_0}} H(P_0))$ is nonsingular, then there exists for ε sufficiently small, solution u_ε of (1.1) with a single peak approaching P_0. The degenerate case was left open.

In [21], Ni and Takagi constructed boundary spike solutions in the case when Ω is axially symmetric. Gui [10] has studied the case when $H(P)$ has a possibly degenerate local maximum at P_0, also constructing multiple-peak solutions at given local maximum points of $H(P)$. In the single peak case, the result in [10] states that for any set $\Lambda \subset \partial\Omega$, open relative to $\partial\Omega$, such that

$$\max_{P\in\Lambda} H(P) > \max_{P\in\partial\Lambda} H(P) \tag{1.2}$$

there exists a family of solutions with a single global maximum point which approaches a local maximum point of $H(P)$ in Λ.

In this paper, we will show that a spike-layer family indeed exists concentrating at any *topologically nontrivial critical point-region*, a variational linking notion first introduced in [5] in the framework of concentration phenomena in nonlinear Schrödinger equations.

This notion includes, for instance, the case of local maxima or local minima of the mean curvature of the boundary, in the same sense as in (1.2), and also that of a possibly degenerate *saddle-point*. More precisely, we can consider a local situation on a set $\Lambda \subset \partial\Omega$ where a change of topology of the level sets of $H(P)$ occurs. If c is the level at which this change takes place in a sense to be made precise below, then a boundary-spike family of solutions exists, with maxima $P_\varepsilon \in \Lambda$ so that $H(P_\varepsilon) \to c$.

Since we do not want to restrict ourselves to the case of a homogeneous nonlinearity, we will consider the more general semilinear Neumann problem

$$\begin{cases} \varepsilon^2 \Delta u - u + f(u) = 0 \quad \text{in } \Omega, \\ u > 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial v} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{1.3}$$

where ε is a small positive number. $f : R \to R$ satisfies the conditions (f1)–(f5) below:

(f1) $f \in C^1(R)$, $f(t) \equiv 0$ for $t \leq 0$, and $f(t) \to +\infty$ as $t \to +\infty$.

(f2) For $t \geq 0$, f admits the decomposition in $C^1(R)$

$$f(t) = f_1(t) - f_2(t),$$

where (i) $f_1(t) \geq 0$, $f_2(t) \geq 0$ with $f_1(0) = f_1'(0) = f_2(0) = f_2'(0) = 0$; and (ii) there is a $q \geq 1$ such that $\frac{f_1(t)}{t^q}$ is nondecreasing in $t > 0$, where as $\frac{f_2(t)}{t^q}$ is nonincreasing in $t > 0$.

(f3) $|f'(t)| \leq a_1 + a_2 t^{p-1}$ for some positive constants a_1, a_2 and $1 < p < (\frac{N+2}{N-2})_+$.

(f4) There exists $\eta \in (0, \frac{1}{2})$ such that $F(t) \leq \eta t f(t)$, $t \geq 0$, where $F(t) = \int_0^t f(s)ds$.

To state the last condition, as in [20], we consider the problem in the whole space

$$\begin{cases} \Delta w - w + f(w) = 0, w > 0 \text{ in } R^N, \\ w(0) = \max\limits_{x\in R^N} w(x) \text{ and } w(x) \to 0 \text{ as } |x| \to +\infty. \end{cases} \tag{1.4}$$

It is well known that (1.4) has a solution w, and w is radial and unique (see [13], [4], [15]). The last condition is stated in (f5).

(f5) $L = \Delta - 1 + f'(w)$ is invertible over $H^2_r(R^N) = \{u \in H^2 : u(x) = u(|x|)\}$.

We note that the function

$$f(t) = t^p - at^q \text{ for } t \geq 0, 1 < q < p$$

with p subcritical and $a \geq 0$ satisfies all the assumptions (see [20]).

Let $H(P)$ be the mean curvature function at $P \in \partial\Omega$. In what follows, we state precisely our assumption on Ω and H. We assume that Ω is a smooth, not necessarily bounded domain in R^N, and that there is an open and bounded set $\Lambda \subset \partial\Omega$ with smooth boundary $\partial\Lambda$ and closed subsets of Λ, B, B_0 such that B is connected and $B_0 \subset B$. Let Γ be the class of all continuous functions $\phi : B \to \Lambda$ with the property that $\phi(y) = y$ for all $y \in B_0$. Assume that the max-min value

$$c = \sup_{\phi \in \Gamma} \min_{y \in B} H(\phi(y)) \tag{1.5}$$

is well defined and additionally that

(H1)

$$\min_{y \in B_0} H(y) > c.$$

(H2) For all $y \in \partial\Lambda$ such that $H(y) = c$, there exists a direction $\hat{T}$, tangent to $\partial\Lambda$ at y so that

$$\nabla H(y) \cdot \hat{T} \neq 0.$$

Note that $\partial\Lambda \subset \partial\Omega$ is an $(N-2)$-dimensional set.

Standard deformation arguments show that these assumptions ensure that the max-min value c is a critical value for $H(P)$ in Λ, which is topologically nontrivial (therefore, our results cover that of [10] in the single peak case). In fact, assumption (H2) "seals" Λ so that the local linking structure described indeed provides critical points at the level c in Λ, possibly admitting full degeneracy.

It is not hard to check that all these assumptions are satisfied in a general local maximum, local minimum, or saddle-point situation, not necessarily nondegenerate or isolated. Our main result asserts that there is a family of solutions to problem (1.1) concentrating around a critical point at the level c of H in Λ.

THEOREM 1.1. *Suppose f satisfies* (f1)–(f5) *and the mean curvature function H satisfied* (H1) *and* (H2). *Then there exists $\varepsilon_0 > 0$ such that when $\varepsilon \leq \varepsilon_0$, problem* (1.3) *has a solution u_ε with the property that*

(i) *u_ε has exactly one local maximum point x_ε and $x_\varepsilon \in \Lambda$;*

(ii) $\lim_{\varepsilon \to 0} H(x_\varepsilon) = c$;

(iii) $\lim_{\varepsilon \to 0} u_\varepsilon(x_\varepsilon + \varepsilon x) = w(x)$ *and there exist positive constants c, δ such that*

$$0 < u_\varepsilon(x) \leq c \, \exp\left(-\frac{\delta|x - x_\varepsilon|}{\varepsilon}\right), \quad x \in \overline{\Omega}.$$

Here w is the unique solution of (1.4).

The proof of this result makes use of ideas developed in [20] and [23] and a variational scheme similar to that in [5], where it is constructed as a bound state for the *semiclassical Schrödinger equation*

$$\varepsilon^2 \Delta u - V(x)u + u^p = 0 \quad \text{in} \quad R^N,$$

exhibiting concentration near topologically nontrivial critical points of $V(x)$; see also the work of the authors in [9]. Related results in this direction can be found in [6] and [7].

We have recently learned that Li [16] has considered, in the case of a bounded domain, a different notion of *nontriviality* not variational in nature. This notion is implied by our assumptions (H1)–(H2) in case the curvature is C^1. Thus, in case $f(s) = u^p$, with p superlinear and subcritical, and for a bounded domain, our result is a consequence of the results in [16]. However, Li's method, relying on a finite-dimensional Lyapunov–Schmidt reduction, is very different from ours.

On the other hand, our method is also applicable to obtain partial localization results even in case H is not C^1.

Finally, we remark that when $p = \frac{N+2}{N-2}$, problem (1.1) has been studied in [1], [2], [3], [11], [18], and [22], among others.

The rest of this paper will be devoted to the proof of Theorem 1.1. In section 2, we define a modified functional which satisfies the Palais–Smale (P.S.) condition and, roughly speaking, permits us to restrict ourselves to what happens in Λ. We then define a min-max value and by using assumption (H1) we prove that there is a critical point for the modified functional with this value. In section 3 by using assumption (H2) we prove that the critical point so found is actually a critical point of the original functional and we conclude the proof of Theorem 1.1.

2. Preliminary results and set-up of a min-max scheme. In this section, we first define a modified functional and state some preliminary results. We then set up a variational scheme and obtain a critical point for the modified functional.

Let $f : R \to R$ satisfying (f1)–(f5). We first define an "energy" functional

$$I_\varepsilon(u) = \frac{1}{2}\int_\Omega \varepsilon^2|\nabla u|^2 + u^2 - \int_\Omega F(u),$$

where $u \in H^1(\Omega)$, $F(u) = \int_0^u f(s)ds$.

As in [5], we now define a modification of this functional which satisfies the P.S. condition and for which we find a critical point via an appropriate min-max scheme.

Let $\mu = \frac{1}{\eta}$, where η is defined by (f4). Let $R > \frac{\mu}{\mu-2}$. Let $a > 0$ be the value at which $f(a)/a = 1/R$. Set

$$\bar{f}(s) = \begin{cases} f(s) & \text{if } s \le a, \\ \dfrac{1}{R}s & \text{if } s > a. \end{cases}$$

The following technical lemma is stated in [10] and can be proved by using local coordinate systems for $\partial\Lambda$.

<u>Lemma 2.1</u>. *There exists a subdomain $\partial\Omega_0 \subset \Omega$ such that $\partial\Omega_0 \cap \partial\Omega = \overline{\Lambda}$ and $\partial\Omega_0^+ := \overline{\partial\Omega_0 \backslash \partial\Omega}$ is smooth and orthogonal to $\partial\Omega$ at $\partial\Lambda$.*

We now define

$$g(\cdot, s) = \chi_{\Omega_0} f(s) + (1-\chi_{\Omega_0})\bar{f}(s) \quad \text{and} \quad G(x,\xi) = \int_0^\xi g(x,\tau)d\tau,$$

where χ_{Ω_0} denotes the characteristic function of Ω_0.

First we note that g is a Carathéodory function. In addition one can check that (f1)–(f4) implies that g satisfies the following conditions:

(g1) $g(x,t) = 0$ for $t \leq 0$ and $g(x,t) \to \infty$ as $t \to \infty$.
(g2) $g(x,t) = o(t)$ near $t = 0$ uniformly in $x \in \Omega$.
(g3) $g(x,t) = O(t^p)$ as $t \to \infty$ for $1 < p < \frac{N+2}{N-2}$ if $N \geq 3$ and no restriction on p if $N = 1, 2$.
(g4) (i) $G(x,t) \leq \mu g(x,t)t \quad \forall x \in \Omega_0, t > 0$

and

(ii) $2G(x,t) \leq g(x,t)t \leq \frac{1}{R}t^2 \quad \forall t \in R^+, x \notin \Omega_0$.

Consider the modified functional

$$J_\varepsilon(u) = \frac{1}{2}\int_\Omega \varepsilon^2|\nabla u|^2 + \frac{1}{2}\int_\Omega u^2 - \int_\Omega G(x,u), \qquad u \in H^1(\Omega),$$

whose critical points correspond to solutions of the equation

$$\begin{cases} \varepsilon^2\Delta u - u + g(u,x) = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial v} = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

As in [5], J_ε satisfies the P.S. condition whether Ω is bounded or not. We observe that a solution to (2.1) which satisfies that $u \leq a$ on $\overline{\Omega}\backslash\Omega_0$ will also be a solution of (1.3). We will define a min-max quantity for J_ε which will yield a solution to (2.1) which turns out to be a solution for (1.3) and thus will be the solution announced by Theorem 1.1.

To this end, we consider the solution manifold of (2.1) defined as

$$M_\varepsilon = \left\{ u \in H^1(\Omega)\backslash\{0\} | \int_\Omega (\varepsilon^2|\nabla u|^2 + u^2) = \int_\Omega g(x,u)u \right\}. \tag{2.2}$$

All nonzero critical points of J_ε of course lie on M_ε. Reciprocally, it is standard to check that critical points of J_ε constrained to this manifold are critical points of J_ε on $H^1(\Omega)$.

Let w be the unique solution of (1.4) and let us consider its energy

$$I(w) = \frac{1}{2}\int_{R^N} (|\nabla w|^2 + w^2) - \int_{R^N} F(w). \tag{2.3}$$

For $P \in \partial\Omega$, we define w_ε^P as

$$w_\varepsilon^P = t_{\varepsilon,P}\, w\left(\frac{x-P}{\varepsilon}\right) \in M_\varepsilon,$$

with $t_{\varepsilon,P} > 0$. Let us consider the center of mass of a function $u \in L^2(\Omega)$ defined as

$$\beta(u) = \frac{\int_{\Omega_0} xu^2 dx}{\int_\Omega u^2 dx}. \tag{2.4}$$

For $P \in B$, it is easy to see that $\beta(w_\varepsilon^P) = P + O(\varepsilon)$. Hence, there exists a continuous function $\tau_\varepsilon(P)$ such that $\tau_\varepsilon(P) = P + O(\varepsilon)$ and $\beta(w_\varepsilon^{\tau_\varepsilon(P)}) = P$ for $P \in B$. We now define

$$w_{\varepsilon,P} = w_\varepsilon^{\tau_\varepsilon(P)}.$$

Hence we have $\beta(w_{\varepsilon,P}) = P \; \forall \; P \in B$, and by similar arguments as in Proposition 3.2 in [19] we find that, $\forall \; P \in B$,

$$J_\varepsilon(w_{\varepsilon,P}) = \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon(N-1)H(P) + o(\varepsilon) \right\}, \tag{2.5}$$

where

$$\gamma := \frac{1}{N+1} \int_{R^N_+} w'(y)^2 y_N dy. \tag{2.6}$$

We now consider the class Γ_ε of all continuous maps $\varphi : B \to M_\varepsilon$ such that

$$\varphi(y) = w_{\varepsilon,y} \qquad \forall y \in B_0,$$

and we define the min-max value S_ε as follows:

$$S_\varepsilon = \inf_{\varphi \in \Gamma_\varepsilon} \sup_{y \in B} J_\varepsilon(\varphi(y)). \tag{2.7}$$

We note that

$$S_\varepsilon \geq \sup_{y \in B_0} J_\varepsilon(w_{\varepsilon,y}) \tag{2.8}$$

and

$$S_\varepsilon = \inf_{\varphi \in \Gamma_\varepsilon} \sup_{y \in B} J_\varepsilon(\varphi(y)) \leq \sup_{y \in B} J_\varepsilon(w_{\varepsilon,y}). \tag{2.9}$$

Hence by (2.5), (2.8), and (2.9), we have

$$\lim_{\varepsilon \to 0} \varepsilon^{-N} S_\varepsilon = \frac{1}{2} I(w). \tag{2.10}$$

The following is the key result of this section. It implies that S_ε is a critical value for J_ε.

LEMMA 2.2. *For ε sufficiently small, we have*

$$S_\varepsilon > \sup_{y \in B_0} J_\varepsilon(w_{\varepsilon,y}). \tag{2.11}$$

In the rest of this section, we prove Lemma 2.2. To this end we will first prove a version of a result of Ni and Takagi for the modified functional J_ε (see Proposition 2.1 in [20]).

LEMMA 2.3. *Let $\Omega_1 \subset \overline{\Omega}$ be a subdomain such that $\partial\Omega_1 \cap \partial\Omega = \Lambda_1$ is open relative to $\partial\Omega$ and $\partial\Omega_1^+ := \overline{\partial\Omega_1 \backslash \partial\Omega}$ is smooth and orthogonal to $\partial\Omega$ at $\partial\Lambda_1$. We define*

$$g_{\Omega_1}(x,u) = \chi_{\Omega_1} f(u) + (1 - \chi_{\Omega_1})\bar{f}(u), \qquad G_{\Omega_1}(x,u) = \int_0^u g_{\Omega_1}(x,s) ds,$$

and

$$J_{\varepsilon,\Omega_1}(u) = \frac{1}{2} \int_\Omega \varepsilon^2 |\nabla u|^2 + \frac{1}{2} \int_\Omega u^2 - \int_\Omega G_{\Omega_1}(x,u).$$

Suppose that u_ε is a solution of

$$\begin{cases} \varepsilon^2 \Delta u - u + g_{\Omega_1}(x,u) = 0 \ \textit{in} \ \Omega, \\ u > 0 \ \textit{in} \ \Omega, \\ \dfrac{\partial u}{\partial v} = 0 \ \textit{on} \ \partial\Omega, \end{cases} \tag{2.12}$$

such that

$$\varepsilon^{-N} J_{\varepsilon,\Omega_1}(u_\varepsilon) \to \frac{1}{2} I(w). \tag{2.13}$$

Then we have

$$J_{\varepsilon,\Omega_1}(u_\varepsilon) = \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon(N-1)H(x_\varepsilon) + o(\varepsilon) \right\}, \tag{2.14}$$

where $x_\varepsilon \in \partial\Omega_1 \cap \partial\Omega$ is the maximum point of u_ε and γ is defined by (2.6). In particular,

$$J_{\varepsilon,\Omega_1}(u_\varepsilon) \geq \varepsilon^N \left\{ \frac{1}{2} I(w) - \varepsilon\gamma \max_{x\in\partial\Omega_1\cap\partial\Omega} (N-1)H(x) + o(\varepsilon) \right\}. \tag{2.15}$$

Before going into the proof of Lemma 2.3 we state and prove a corollary that will be useful later.

COROLLARY 2.1. *Let $\varepsilon = \varepsilon_k \to 0$ and $u_\varepsilon \in M_{\varepsilon,\Omega_1}$ be a family of functions such that*

$$\limsup_{\varepsilon\to 0} \varepsilon^{-N} J_{\varepsilon,\Omega_1}(u_\varepsilon) \leq \frac{1}{2} I(w), \tag{2.16}$$

where

$$M_{\varepsilon,\Omega_1} = \left\{ u \in H^1(\Omega)\backslash\{0\} | \int_\Omega (\varepsilon^2|\nabla u|^2 + u^2) = \int_\Omega g_{\Omega_1}(x,u)u \right\}.$$

Let $x_\varepsilon = \beta(u_\varepsilon)$ be the center of mass of u_ε; then $x_\varepsilon \to \partial\Omega$, and if $\bar{x}$ is an accumulation point of $\{x_\varepsilon\}$, the following estimate holds:

$$J_{\varepsilon,\Omega_1}(u_\varepsilon) \geq \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon(N-1)H(\bar{x}) + o(\varepsilon) \right\}. \tag{2.17}$$

Proof. Passing to a subsequence we can assume that $x_\varepsilon \to \bar{x}$. Let us consider the modified center of mass defined as

$$\bar{\beta}(u) = \frac{\int_{B_\delta(\bar{x})} xu^2}{\varepsilon^N \int_{R^N} w^2}.$$

Given $\delta > 0$ we then have that

$$\bar{\beta}(u_\varepsilon) \in B_\delta(\bar{x}) \tag{2.18}$$

$\forall$ small ε. In fact, using a concentration-compactness-type argument similar to the one given in Lemma 1.1 in [5], we find $R > 0$, a subsequence $\varepsilon \to 0$, and $y_\varepsilon \in \Omega_\varepsilon = \varepsilon^{-1}\Omega$ such that

$$\int_{B_R(y_\varepsilon)} v_\varepsilon^2 \geq \sigma > 0,$$

where $v_\varepsilon(x) = v_\varepsilon(\varepsilon x)$.

Let us assume first that $dist(y_\varepsilon, \partial\Omega_\varepsilon) \to \infty$. Since v_ε is bounded in $H^1(\Omega_\varepsilon)$, given $\delta > 0$ there exists $r > 0$ such that

$$\int_{B_{r+1}(0)\setminus B_r(0)} |\nabla u_\varepsilon|^2 + u_\varepsilon^2 \leq \delta.$$

Then we choose an appropriate cut-off function ψ so that $\psi = 1$ on $B_r(0)$ and $\psi = 0$ on $B_{r+1}(0)$ and we find

$$u_\varepsilon = \psi u_\varepsilon + (1-\psi)u_\varepsilon = w_\varepsilon + v_\varepsilon.$$

If we choose δ small enough, we find that for both v_ε and w_ε we can find t_ε^1, t_ε^2 very close to 1 so that $\tilde{w}_\varepsilon = t_\varepsilon^1 w_\varepsilon$ and $\tilde{v}_\varepsilon = t_\varepsilon^2 v_\varepsilon$ are in M_{ε,Ω_1}. But this implies that lim inf $J_{\varepsilon,\Omega_1}(u_\varepsilon) \geq I(w)$, contradicting the hypothesis.

Therefore, we must have that $dist(y_\varepsilon, \partial\Omega_\varepsilon) \leq C$. We can assume that $y_\varepsilon \in \partial\Omega_\varepsilon$. By the argument given above, taking a sequence $\delta_n \to 0$ and using (2.16) we find a subsequence $u_\varepsilon = v_\varepsilon + w_\varepsilon$ with $w_\varepsilon \to 0$.

Finally, using the minimizing character of this sequence u_ε and Ekeland's variational principle we find that $u_\varepsilon(x_\varepsilon + \varepsilon y)$ converges in H^1-sense to a least energy critical point w of the limiting functional I given in (2.3) in the half space. We certainly have that $x_\varepsilon + \varepsilon y_\varepsilon \to x \in \partial\Omega$, thus proving (2.18).

Then we have

$$J_{\varepsilon,\Omega_1}(u_\varepsilon) \geq \inf\{J_{\varepsilon,\Omega_1}(u) \mid u \in M_{\varepsilon,\Omega_1},\ \bar{\beta}(u) \in B_\delta(\bar{x})\}.$$

Since the functional J_{ε,Ω_1} satisfies the P.S. condition, it follows that the latter number is attained at some function $\bar{u}_\varepsilon \in H^1(\Omega)$. Working out a first variation with test functions supported outside $B_\delta(\bar{x})$, we see that $\bar{u}_\varepsilon$ satisfies the equation

$$\varepsilon^2\Delta\bar{u}_\varepsilon - \bar{u}_\varepsilon + g_{\Omega_1}(x, \bar{u}_\varepsilon) = 0 \text{ in } \Omega\setminus B_\delta(\bar{x}).$$

Again, if we set $v_\varepsilon(y) = \bar{u}_\varepsilon(\bar{x}_\varepsilon + \varepsilon y)$ with $\bar{x}_\varepsilon = \beta(\bar{u}_\varepsilon)$, then v_ε converges in the H^1-sense to w in the half space. In particular, elliptic estimates applied to the above equation imply that $\bar{u}_\varepsilon$ goes to zero uniformly, away from the ball $B_\delta(\bar{x})$. Thus we have that

$$J_{\varepsilon,\Omega_1}(\bar{u}_\varepsilon) = J_{\varepsilon,\Omega_1\cap B_{2\delta}(\bar{x})}(\bar{u}_\varepsilon)$$

and also $\bar{u}_\varepsilon \in M_{\varepsilon,\Omega_1\cap B_{2\delta}(\bar{x})}$. Let us consider a set Ω_δ so that $\Omega_1 \cap B_{2\delta}(\bar{x}) \subset \Omega_\delta \subset \Omega_1 \cap B_{3\delta}(\bar{x})$, satisfying the hypotheses of Lemma 2.3. Then we obtain

$$J_{\varepsilon,\Omega_1}(\bar{u}_\varepsilon) \geq \inf_{u \in M_{\varepsilon,\Omega_\delta}} J_{\varepsilon,\Omega_\delta}(u).$$

However the latter number can be estimated from below using Lemma 2.3. Doing so we have

$$J_{\varepsilon,\Omega_1}(\bar{u}_\varepsilon) \geq \varepsilon^N \left\{\frac{1}{2}I(w) - \varepsilon\gamma \max_{x\in\partial\Omega_\delta\cap\partial\Omega}(N-1)H(x) + o(\varepsilon)\right\}.$$

To obtain (2.17), we first use the continuity of H to choose δ and then we choose ε small enough, according to (2.15). This finishes the proof. $\square$

Now we will give a proof of Lemma 2.3. We start with some preliminaries.

Proof of Lemma 2.3. Since u_ε satisfies (2.12) and $\varepsilon^{-N} J_{\varepsilon,\Omega_1}(u_\varepsilon)$ is bounded, u_ε converges locally in the H^1 sense to a solution of the limiting equation. Then a concentration-compactness argument gives that $\|\tilde{u}_\varepsilon - w\|_{H^1(\Omega_{\varepsilon,z_\varepsilon})} \to 0$ for some $z_\varepsilon \in \overline{\Omega}$, where

$$\Omega_{\varepsilon,P} = \{y | \varepsilon y + P \in \overline{\Omega}\}, \qquad P \in \overline{\Omega},$$

and $\tilde{u}_\varepsilon(y) = u_\varepsilon(\varepsilon y + z_\varepsilon)$. Moreover, because of (2.13) we have that $\frac{d(z_\varepsilon,\partial\Omega)}{\varepsilon} \leq C$ and $z_\varepsilon \in \Omega_1$ (otherwise, the energy of u_ε will be at least of the order of $\varepsilon^N I(w)$; see Lemma 1.1 in [5]). Observe that u_ε satisfies

$$\varepsilon^2 \Delta u_\varepsilon - u_\varepsilon + f(u_\varepsilon) + h_\varepsilon = 0, \tag{2.19}$$

where $h_\varepsilon = (1 - \chi_{\Omega_1})(\bar{f}(u_\varepsilon) - f(u_\varepsilon))$. Hence $h_\varepsilon = o(1)$ uniformly and $\tilde{u}_\varepsilon \to w$ in a C^1_{loc} sense. Furthermore, there exist constants α, $\beta > 0$ such that

$$\tilde{u}_\varepsilon(y) \leq \alpha \ \exp(-\beta|y|).$$

Next, an argument given in [19] shows that u_ε has only one local maximum point x_ε and $x_\varepsilon \in \partial\Omega_1 \cap \partial\Omega$.

We now consider two cases. Let $b > 0$ so that $w(b) = a$.

Case 1. If lim $\inf_{\varepsilon\to 0} d(x_\varepsilon, \partial\Omega_1^+)/\varepsilon > b$, then u_ε satisfies

$$\varepsilon^2 \Delta u_\varepsilon - u_\varepsilon + f(u_\varepsilon) = 0,$$

and then, by Proposition 2.1 in [20], we have that

$$J_{\varepsilon,\Omega_1}(u_\varepsilon) = \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon(N-1)H(x_\varepsilon) + o(\varepsilon) \right\},$$

finishing the proof of the lemma.

Case 2. lim $\inf_{\varepsilon\to 0} d(x_\varepsilon, \partial\Omega_1^+)/\varepsilon \leq b$. We see first that we can assume that lim $\inf_{\varepsilon\to 0} d(x_\varepsilon, \partial\Omega_1^+)/\varepsilon = b$ since the contrary, together with the convergence of $\tilde{u}_\varepsilon$ to w, implies a contradiction with (2.13).

To prove the lemma in this case we need some work. We next consider some notation. Let $\bar{x}_\varepsilon \in \partial\Omega_1^+$ be such that $d(x_\varepsilon, \partial\Omega_1^+) = |x_\varepsilon - \bar{x}_\varepsilon|$. Then since $\partial\Omega_1^+$ is orthogonal to $\partial\Omega$ at Λ_1, we have that the projection of $\bar{x}_\varepsilon$ onto $\partial\Lambda_1$, which we call $\bar{x}_\varepsilon^p$, satisfies

$$\frac{|x_\varepsilon - \bar{x}_\varepsilon^p|}{\varepsilon} \to b \qquad \text{and} \qquad \frac{|\bar{x}_\varepsilon - \bar{x}_\varepsilon^p|}{\varepsilon} \to 0. \tag{2.20}$$

Without loss of generality, we can assume that $\nu_{x_\varepsilon} = -e_N$, where ν_{x_ε} denotes the exterior normal at x_ε and that $\bar{x}_\varepsilon = d(x_\varepsilon, \partial\Omega_1^+)e_1^\varepsilon$, where $e_1^\varepsilon \to e_1$ as $\varepsilon \to 0$.

Set $x = x_\varepsilon + \varepsilon y$, $\Omega_\varepsilon = \{y : x_\varepsilon + \varepsilon y \in \Omega\}$. For notational convenience in the rest of the paper, given a function $p : \Omega \to R$, we denote by $\tilde{p}$ the function defined on Ω_ε as $\tilde{p}(y) = p(x)$. We observe that support of the function $\tilde{h}_\varepsilon$ is contained in $B_{\delta_\varepsilon}((\bar{x}_\varepsilon - x_\varepsilon)/\varepsilon) \cap \overline{\Omega}_\varepsilon$, where $\delta_\varepsilon \to 0$. This fact follows from the uniform convergence of $\tilde{u}_\varepsilon$ to w and the exponential decay of w at infinity.

Now we will study the asymptotic behavior of u_ε. First we define the function ϕ_ε as

$$u_\varepsilon(x) = w_\varepsilon(x) + \varepsilon\phi_\varepsilon, \qquad x \in \Omega, \tag{2.21}$$

where $w_\varepsilon(x) = w(\frac{x-x_\varepsilon}{\varepsilon})$. It is our goal to study the behavior of the function ϕ_ε. The next lemma provides an important estimate.

LEMMA 2.4. *For ε sufficiently small, we have*

$$\|\tilde{h}_\varepsilon\|_{L^1(\Omega_\varepsilon)} \leq o(\varepsilon). \tag{2.22}$$

Proof. We multiply the equation satisfied by $\tilde{u}_\varepsilon$ (see (2.19)) by $\frac{\partial \tilde{u}_\varepsilon}{\partial y_1}$ and integrate by parts to obtain

$$\int_{\Omega_\varepsilon} \tilde{h}_\varepsilon \frac{\partial \tilde{u}_\varepsilon}{\partial y_1} = \int_{\partial\Omega_\varepsilon} \left\{ F(\tilde{u}_\varepsilon) - \frac{1}{2}\tilde{u}_\varepsilon^2 \right\} \nu_1 dy,$$

where ν_1 is the first component of the normal vector. To estimate the right-hand side of the above equality we give a local representation of the boundary near the origin and find that $\nu_1 = \varepsilon \sum_{i=1}^{N-1} \alpha_i y_i + O(\varepsilon^2)$. On the other hand, from the radial symmetry of w we have that

$$\int_{\partial R^N_+} \left\{ F(w) - \frac{1}{2} w^2 \right\} y_i dy = 0 \qquad \text{for} \quad i = 1, \ldots, N-1. \tag{2.23}$$

Then

$$\int_{\partial\Omega_\varepsilon} \left\{ F(\tilde{u}_\varepsilon) - \frac{1}{2}\tilde{u}_\varepsilon^2 \right\} \nu_1 dy = o(\varepsilon). \tag{2.24}$$

To finish we observe that since $\text{supp}(\tilde{h}_\varepsilon) \subset B_{2\delta_\varepsilon}(be_1)$, for small ε, we have that $\frac{\partial \tilde{u}_\varepsilon}{\partial u_1} \to \frac{\partial w}{\partial y_1}(be_1) \neq 0$ for all $y \in \text{supp}(\tilde{h})$ and hence

$$\int_{\Omega_\varepsilon} \tilde{h}_\varepsilon = o(\varepsilon),$$

proving (2.22). □

Next we study the behavior of the function $\tilde{\phi}_\varepsilon$. We see that $\tilde{\phi}_\varepsilon$ satisfies the equation

$$\begin{cases} \Delta\tilde{\phi}_\varepsilon - (1 + d_\varepsilon)\tilde{\phi}_\varepsilon + f'(w)\tilde{\phi}_\varepsilon + \dfrac{\tilde{h}_\varepsilon}{\varepsilon} = 0 \text{ in } \Omega_\varepsilon, \\ \dfrac{\partial\tilde{\phi}_\varepsilon}{\partial\nu} = -\dfrac{1}{\varepsilon}\dfrac{\partial w}{\partial \nu} \text{ on } \partial\Omega_\varepsilon, \end{cases} \tag{2.25}$$

where

$$d_\varepsilon = \frac{1}{\varepsilon\tilde{\phi}_\varepsilon}(f(\tilde{u}_\varepsilon) - f(w)) - f'(w).$$

We observe that $d_\varepsilon \to 0$ uniformly and we note that $\tilde{w}_\varepsilon = w$.

A local representation of Ω near x_ε is considered next. There is $R > 0$ and a neighborhood N_ε of x_ε so that $(y', y_N) \in N_\varepsilon \cap \Omega$ if and only if $y_N > \rho_\varepsilon(y')$,

$y' \in B(0,R)$, $\rho_\varepsilon(0) = x_\varepsilon$, and $\nabla\rho_\varepsilon(0) = 0$. We observe that if $x_\varepsilon \to x_0$ as $\varepsilon \to 0$, then $\rho_\varepsilon \to \rho$ in C^3 uniformly, where ρ is a local representation of the boundary centered at x_0.

Now we get an asymptotic formula for the normal derivative of w. We find, for $y \in B(0, \frac{R}{\varepsilon})$, that

$$\frac{\partial w}{\partial \nu}(y, \tilde{\rho}_\varepsilon(y)) = \frac{\varepsilon w'(|y|)}{2|y|}(\rho_\varepsilon)_{ij} y_i y_j + o(\varepsilon), \tag{2.26}$$

where $(\rho_\varepsilon)_{ij}$ denotes the partial derivatives of ρ_ε at 0. Here and in what follows we use the Einstein convention for summations.

In studying the behavior of $\tilde{\phi}_\varepsilon$ we need the limiting equation

$$\begin{cases} \Delta\phi - \phi + f'(w)\phi = 0 \text{ in } R^N_+, \\ \dfrac{\partial\hat{\phi}}{\partial y_N} = -\dfrac{w'(|y|)}{2|y|}\rho_{ij} y_i y_j \quad \text{on } \partial R^N_+. \end{cases} \tag{2.27}$$

We have the following lemma.

Lemma 2.5. *There is $1 < q < N/(N-1)$ so that $\|\tilde{\phi}_\varepsilon\|_{L^q(\Omega_\varepsilon)}$ is bounded and there are constants α, β, $R_0 > 0$ so that*

$$|\tilde{\phi}_\varepsilon(y)| \le \alpha \ \exp(-\beta|y|) \qquad \textit{for} \quad |y| > R_0. \tag{2.28}$$

Moreover,

$$\|\tilde{\phi}_\varepsilon - \tilde{\phi}_0\|_{L^q(\Omega_\varepsilon)} \to 0, \tag{2.29}$$

where $\tilde{\phi}_0 \in H^1(R^N_+)$ is the solution to (2.27).

Proof. Let us assume that $\|\tilde{\phi}_\varepsilon\|_{L^q(\Omega_\varepsilon)}$ is not bounded and define the function $\hat{\phi}_\varepsilon = \tilde{\phi}_\varepsilon / \|\tilde{\phi}_\varepsilon\|_{L^q(\Omega)_\varepsilon)}$. Then $\hat{\phi}_\varepsilon$ satisfies

$$\begin{cases} \Delta\hat{\phi}_\varepsilon - (1+d_\varepsilon)\hat{\phi}_\varepsilon + f'(w)\hat{\phi}_\varepsilon + \hat{h}_\varepsilon = 0 \text{ in } \Omega_\varepsilon, \\ \dfrac{\partial\hat{\phi}_\varepsilon}{\partial\nu} = n_\varepsilon \text{ on } \partial\Omega_\varepsilon, \end{cases} \tag{2.30}$$

where $\hat{h}_\varepsilon = \tilde{h}_\varepsilon / \varepsilon \|\tilde{\phi}_\varepsilon\|_{L^q(\Omega_\varepsilon)} \to 0$ in the L^1 sense and

$$n_\varepsilon = -\frac{1}{\varepsilon}\frac{\partial w}{\partial\nu} / \|\tilde{\phi}_\varepsilon\|_{L^q(\Omega_\varepsilon)}.$$

We observe that $n_\varepsilon \to 0$ uniformly and that it satisfies an estimate of the form

$$|n_\varepsilon(y)| \le \alpha_\varepsilon \ \exp(-\bar{\beta}|y|) \qquad \text{for} \quad y \in \partial\Omega_\varepsilon \tag{2.31}$$

for some constants α_ε, $\bar{\beta} > 0$, and $\alpha_\varepsilon \to 0$.

We recall that $\text{supp}(\tilde{h}_\varepsilon) \subset B_{2\delta_\varepsilon}(be_1)$, with $\delta_\varepsilon \to 0$. Thus, standard elliptic estimates and comparison arguments, using the facts just mentioned and that $\|\hat{h}_\varepsilon\|_{L^q(\Omega_\varepsilon)}$ is bounded, yield the existence of constants R_0, α, $\beta > 0$ such that

$$|\hat{\phi}_\varepsilon(y)| \le \alpha \ \exp(-\beta|y|) \qquad \text{for} \quad |y| > R_0. \tag{2.32}$$

Since $\|\Delta\hat{\phi}_\varepsilon\|_{L^1(\Omega_\varepsilon)} \le C$, a well-known elliptic estimate yields that

$$\|\hat{\phi}_\varepsilon\|_{W^{1,q}(\Omega_\varepsilon \cap B_{R_0}(0))} \le C_{R_0}. \tag{2.33}$$

By the boundedness of $\hat{\phi}_\varepsilon$ in L^q we have that for a subsequence $\hat{\phi}_\varepsilon \rightharpoonup \hat{\phi}$ weakly in L^q. Now, (2.32) and (2.33) implies that this convergence is strong in L^q, in particular, $\hat{\phi} \ne 0$. Moreover, $\hat{\phi} \in W^{1,q}(R^N_+)$, it satisfies

$$\begin{cases} \Delta\hat{\phi} - \hat{\phi} + f'(w)\hat{\phi} = 0 \text{ in } R^N_+, \\ \dfrac{\partial \hat{\phi}}{\partial y_N} = 0 \text{ on } \partial R^N_+ \end{cases} \tag{2.34}$$

and

$$|\hat{\phi}(y)| \le \alpha\ \exp(-\beta|y|) \qquad \text{for} \quad |y| \quad \text{large}. \tag{2.35}$$

We observe that $\nabla w(0) = 0$ and that $\nabla u_\varepsilon(x_\varepsilon) = 0$; then $\nabla\hat{\phi}_\varepsilon(0) \to \nabla\hat{\phi}(0) = 0$. Thus hypothesis (f5) and the argument given in the proof of Lemma 4.6 of Ni and Takagi [20] imply that $\hat{\phi} \equiv 0$, which is a contradiction.

Next we can give a similar argument to obtain that the family $\tilde{\phi}_\varepsilon$ satisfies (2.28) and that, if $\tilde{\phi}_0$ is the solution of (2.27), then

$$\|\tilde{\phi}_\varepsilon - \tilde{\phi}_0\|_{L^q(\Omega_\varepsilon)} \to 0, \tag{2.36}$$

finishing the proof of Lemma 2.5. □

Proof of Lemma 2.3 (*finished*). We have

$$\begin{aligned} \varepsilon^{-N} J_{\varepsilon,\Omega_1}(u_\varepsilon) &= \int_{\Omega_\varepsilon} \frac{1}{2}(|\nabla \tilde{u}_\varepsilon|^2 + \tilde{u}_\varepsilon^2) - F(\tilde{u}_\varepsilon) - F(\tilde{u}_\varepsilon) + \int_{\Omega_\varepsilon \setminus \Omega_{1\varepsilon}} \bar{F}(\tilde{u}_\varepsilon) - F(\tilde{u}_\varepsilon). \\ &= I_1 + I_2 \end{aligned}$$

We first estimate integral I_2. It follows from hypothesis (f5) and Lemma 2.4 that

$$\begin{aligned} |I_2| &= \int_{\Omega_\varepsilon} (1 - \chi_{\Omega_{1\varepsilon}})(F(\tilde{u}_\varepsilon) - \bar{F}(\tilde{u}_\varepsilon)) \\ &= \int_{\Omega_\varepsilon} (1 - \chi_{\Omega_{1\varepsilon}}) \int_0^{\tilde{u}_\varepsilon} (f(s) - \bar{f}(s))ds \\ &\le \int_{\Omega_\varepsilon} (1 - \chi_{\Omega_{1\varepsilon}}) \frac{f(\tilde{u}_\varepsilon) - \bar{f}(\tilde{u}_\varepsilon)}{\tilde{u}_\varepsilon} \frac{\tilde{u}_\varepsilon^2}{2} = o(\varepsilon). \end{aligned} \tag{2.37}$$

Next we study I_1; for that purpose, we write

$$\begin{aligned} I_1 &= \int_{\Omega_\varepsilon} \frac{1}{2}(|\nabla w|^2 + w^2) - F(w) + \\ &+ \varepsilon \int_{\Omega_\varepsilon} \{\nabla w \cdot \nabla\tilde{\phi}_\varepsilon + w\tilde{\phi}_\varepsilon - f(w)\tilde{\phi}_\varepsilon\} + E_\varepsilon = I_1' + I_2' + E_\varepsilon. \end{aligned} \tag{2.38}$$

A direct computation using the properties of w yields

$$I_1' = \frac{1}{2} I(w) - \gamma\varepsilon(N-1)H(x_\varepsilon) + o(\varepsilon). \tag{2.39}$$

Using integration by parts and the equation satisfied by w we find

$$I_2' = \varepsilon \int_{\partial\Omega_\varepsilon} \frac{\partial w}{\partial \nu} \tilde{\phi}_\varepsilon = o(\varepsilon), \tag{2.40}$$

where the last equality follows from (2.26) and the fact that $\varepsilon\tilde{\phi}_\varepsilon \to 0$ uniformly. Finally we consider E_ε: using Taylor expansion we have

$$E_\varepsilon = \varepsilon^2 \left\{ \int_0^1 (1-t) \left(\int_{\Omega_\varepsilon} |\nabla \tilde{\phi}_\varepsilon|^2 + \tilde{\phi}_\varepsilon^2 - f'(w + t\varepsilon\tilde{\phi}_\varepsilon)\tilde{\phi}_\varepsilon^2 \right) dt \right\}. \tag{2.41}$$

For a given large R, we obtain

$$\begin{aligned} \varepsilon \int_{\Omega_\varepsilon} |\nabla \tilde{\phi}_\varepsilon|^2 = {} & \varepsilon \int_{\Omega_\varepsilon \cap B_R(0)} |\nabla \tilde{\phi}_\varepsilon|^2 + \varepsilon \int_{\partial(\Omega_\varepsilon \cap B_R(0))} \nabla \tilde{\phi}_\varepsilon \cdot \nu \tilde{\phi}_\varepsilon \\ & - \varepsilon \int_{\Omega_\varepsilon \cap B_R(0)^c} \Delta \tilde{\phi}_\varepsilon \tilde{\phi}_\varepsilon. \end{aligned} \tag{2.42}$$

The first and second term on the right-hand side above go to zero because $\varepsilon\tilde{\phi}_\varepsilon \to 0$ in C^1_{loc} and $\tilde{\phi}_\varepsilon \in W^{1,q}_{loc}(\Omega_\varepsilon)$. Next, using the equation for $\tilde{\phi}_\varepsilon$ and (2.28) we find that the third term also converges to 0, so we conclude that

$$\varepsilon \int_{\Omega_\varepsilon} |\nabla \tilde{\phi}_\varepsilon|^2 = o(1). \tag{2.43}$$

Using similar arguments we treat the other terms appearing in (2.41). Thus we finally obtain that $E_\varepsilon = o(\varepsilon)$, finishing the proof. □

Proof of Lemma 2.2. Suppose (2.11) is not true; then

$$S_\varepsilon = \sup_{y \in B_0} J_\varepsilon(w_{\varepsilon,y}). \tag{2.44}$$

Hence

$$\begin{aligned} S_\varepsilon &= \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon \min_{y \in B_0} (N-1) H(y) + o(\varepsilon) \right\} \\ &\leq \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon(c+\delta) + o(\varepsilon) \right\}, \end{aligned}$$

where $c+\delta \leq \min_{y\in B_0} H(y)$ for some $\delta > 0$ (by assumption (H1)). Then, by definition of S_ε there exists $\varphi_\varepsilon \in \Gamma_\varepsilon$ such that

$$J_\varepsilon(\varphi_\varepsilon(y)) \leq \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon \left(c + \frac{\delta}{2} \right) + o(\varepsilon) \right\} \quad \forall\, y \in B. \tag{2.45}$$

Take a sequence $\varepsilon_n \to 0$ and denote $\varphi_{\varepsilon_n} = \varphi_n$. Let Λ^+ be a small fixed neighborhood of Λ and $\pi : \Lambda^+ \to \Lambda$ a continuous map which equals the identity on Λ. Define $\phi_n(y) = \pi(\beta(\varphi_n(y)))$ for $y \in B$, where β is the center of mass defined in (2.4). We claim that for large n we have

$$\beta(\varphi_n(y)) \in \Lambda^+ \quad \text{and} \quad H(\phi_n(y)) \geq c + \frac{\delta}{4} \quad \forall\, y \in B. \tag{2.46}$$

This immediately yields the desired contradiction. In fact, since $\varphi_n(y) = w_{\varepsilon_n,y}$ for $y \in B_0$, it follows that $\phi_n(y) = y$ for $y \in B_0$. Hence $\phi_n \in \Gamma$ and by definition of c, we have

$$c \geq \min_{y\in B} H(\phi_n(y)), \tag{2.47}$$

which is impossible in view of (2.46).

We now prove (2.46). The fact that $\beta(\varphi_n(y)) \in \Lambda^+$ is obtained by slightly modifying the arguments in [5, Lemma 1.1]. Thus, we just need to prove the second statement in (2.46). Suppose it is not true; then there exists $y_n \in B$ such that

$$H(\phi_n(y_n)) \leq c + \frac{\delta}{4}.$$

We can assume that $\phi_n(y_n) \to x_0 \in \bar{\Lambda}$ and then $H(x_0) \leq c + \frac{\delta}{4}$.

Next we apply Corollary 2.1 to the family of functions $u_n = \varphi_n(y_n)$ and obtain that

$$J_\varepsilon(u_n) \geq \varepsilon_n^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon \left(c + \frac{\delta}{4} \right) + o(\varepsilon) \right\}. \tag{2.48}$$

Comparing (2.45) and (2.48) we get a contradiction and thus Lemma 2.2 is proved. □

By Lemma 2.2, we have by a standard deformation argument the main result of this section, namely, the following proposition.

PROPOSITION 2.6. *The number defined by* (2.8) *is a critical value of* J_ε. *That is, there is a solution* $u_\varepsilon \in H^1$ *to* (2.1) *such that* $J_\varepsilon(u_\varepsilon) = S_\varepsilon$ $\forall$ ε *sufficiently small.*

3. Proof of Theorem 1.1. In this section, we show that the solution u_ε to (2.1) constructed in Proposition 2.6 is a solution of (1.3). The key step is the following proposition.

PROPOSITION 3.1. *If* m_ε *is given by* $m_\varepsilon = \max_{x\in\partial\Omega_0} u_\varepsilon(x)$, *then*

$$\lim_{\varepsilon\to 0} m_\varepsilon = 0. \tag{3.1}$$

Before we prove the above proposition, we need the following lemma.

LEMMA 3.2. *Let* x_ε *be the maximum point of* u_ε; *then we have*

$$\lim_{\varepsilon\to 0} H(x_\varepsilon) \to c,$$

where c *is the max-min value defined in* (1.5).

Proof. By Lemma 2.3, we have

$$J_\varepsilon(u_\varepsilon) = \varepsilon^N \left\{ \frac{1}{2} I(w) - \gamma\varepsilon(N-1)H(x_\varepsilon) + o(\varepsilon) \right\} \tag{3.2}$$

and then

$$\limsup_{\varepsilon\to 0} H(x_\varepsilon) \leq c. \tag{3.3}$$

In fact, assuming the contrary we have $H(x_\varepsilon) \geq c + \frac{\delta}{2}$ for ε and δ small and then we have a similar situation as in (2.45), so that following the arguments given from there we get a contradiction.

On the other hand, let $\delta > 0$ and $\phi_0 \in \Gamma$ be such that

$$\min_{y \in B} H(\phi_0(y)) \geq c - \delta.$$

Then, by (2.5) and the definition of $S_\varepsilon = J_\varepsilon(u_\varepsilon)$, we have

$$\begin{aligned} J_\varepsilon(u_\varepsilon) &\leq \sup_{y \in B} J_\varepsilon(w_{\varepsilon,\phi_0(y)}) \\ &\leq \varepsilon^n \left\{ \frac{1}{2} I(w) - \varepsilon\gamma(N-1) \min_{y \in B} H(\phi_0(y)) + o(\varepsilon) \right\} \\ &\leq \varepsilon^n \left\{ \frac{1}{2} I(w) - \varepsilon\gamma(N-1)(c - \delta) + o(\varepsilon) \right\}. \end{aligned} \tag{3.4}$$

From here and (3.2) we obtain

$$H(x_\varepsilon) \geq c - \delta + o(1).$$

Since δ is arbitrary using (3.3) we then conclude with the proof. □

We are now in a position to prove Proposition 3.1.

Proof of Proposition 3.1. Suppose, on the contrary, that $m_\varepsilon \geq \delta > 0$. Then let $u_\varepsilon(x_\varepsilon) = \max_{x \in \bar{\Omega}} u_\varepsilon(x)$. Then $x_\varepsilon \in \overline{\Lambda}$, $\frac{d(x_\varepsilon, \partial\Lambda)}{\varepsilon} \to b > 0$, and $w(b) = a$, and by Lemma 3.2 $H(x_\varepsilon) \to c$ as $\varepsilon \to 0$. We recall that the function $\tilde{u}_\varepsilon$ satisfies

$$\begin{cases} \Delta\tilde{u}_\varepsilon - \tilde{u}_\varepsilon + f(\tilde{u}_\varepsilon) + \tilde{h}_\varepsilon = 0 \text{ in } \Omega_\varepsilon, \\ \dfrac{\partial \tilde{u}_\varepsilon}{\partial \nu} = 0 \text{ on } \partial\Omega_\varepsilon. \end{cases} \tag{3.5}$$

Let $\hat{T}_\varepsilon$ be a direction, tangent to Λ_ε at $\bar{x}_\varepsilon^p$. We assume that $\hat{T}_\varepsilon$ converges to $\hat{T}_0$ and we observe that $\hat{T}_0 \perp e_N$, with the notational convention given in the proof of Lemma 2.3. Next we multiply (3.5) by $\nabla\tilde{u}_\varepsilon \cdot \hat{T}_\varepsilon$ and we integrate by parts to obtain

$$\int_{\partial\Omega_\varepsilon} \left\{ \frac{|\nabla\tilde{u}_\varepsilon|^2}{2} + \frac{\tilde{u}_\varepsilon^2}{2} - F(\tilde{u}_\varepsilon) \right\} \hat{T}_\varepsilon \cdot \nu = \int_{\Omega_\varepsilon} \tilde{h}_\varepsilon \frac{\partial \tilde{u}_\varepsilon}{\partial \hat{T}_\varepsilon}. \tag{3.6}$$

Using the asymptotic expansion (2.21), integrating by parts again, and using the equation for w we obtain that

$$\begin{aligned} \int_{\partial\Omega_\varepsilon} \frac{\partial w}{\partial \nu} \frac{\partial w}{\partial \hat{T}_\varepsilon} + \varepsilon \int_{\partial\Omega_\varepsilon} \int_0^1 \left\{ \nabla\tilde{u}_\varepsilon(t) \cdot \nabla\tilde{\phi}_\varepsilon + \tilde{u}_\varepsilon(t)\tilde{\phi}_\varepsilon - f(\tilde{u}_\varepsilon(t))\tilde{\phi}_\varepsilon \right\} \hat{T}_\varepsilon \cdot \nu \, dt \\ = \int_{\Omega_\varepsilon} \tilde{h}_\varepsilon \frac{\partial \tilde{u}_\varepsilon}{\partial \hat{T}_\varepsilon}, \end{aligned} \tag{3.7}$$

where $\tilde{u}_\varepsilon(t) = w + t\varepsilon\tilde{\phi}_\varepsilon$. For later reference, we write $I_1 + I_2 = I_3$ above. We first claim that by slightly modifying $\hat{T}_\varepsilon$ we can get $I_3 = 0$. In fact, the normal vector ν near the origin, in a ball of fixed radius $R_0 > 0$, has the form

$$\nu = 0(1 + O(\varepsilon))e_N + \varepsilon \sum_{i=1}^{N-1} \vec{\alpha}_i y_i + o(\varepsilon). \tag{3.8}$$

Thus, taking into account that the support of $\tilde{h}_\varepsilon$ shrinks to a point, that $\tilde{h}_\varepsilon \geq 0$, and that $\tilde{u}_\varepsilon$ converges to w, we perturb $\hat{T}_\varepsilon$ so that $\hat{T}_\varepsilon \perp e_N$ and $I_3 = 0$, and still keep that $\hat{T}_\varepsilon \to \hat{T}_0$.

Next we consider I_2. We observe that

$$\int_{\partial R^N_+} \left\{ \nabla w \cdot \nabla \tilde{\phi}_0 + w\tilde{\phi}_0 - f(w)\tilde{\phi}_0 \right\} y_i = 0, \tag{3.9}$$

since the function $\tilde{\phi}_0$, the solution of (2.27), is even on the boundary and so is w. From here, and taking into account (3.8), (2.28), and the convergence of $\tilde{\phi}_\varepsilon$ to ϕ_0 in $W^{1,q}_{loc}(\Omega_\varepsilon)$, we find that $I_2 = o(\varepsilon^2)$ and thus

$$\int_{\partial\Omega_\varepsilon} \frac{\partial w}{\partial \nu} \frac{\partial w}{\partial \hat{T}_\varepsilon} = o(\varepsilon^2). \tag{3.10}$$

Now we turn to study this last term. For this purpose, we obtain an expansion for derivatives of w near the origin and on the boundary of Ω_ε. A direct calculation, using Taylor expansion of the function w and the local representation the boundary, with the notation given in section 2, gives

$$w_\ell(y, \tilde{\rho}_\varepsilon(y)) = w_\ell(y, 0) + O(\varepsilon^2), \qquad 1 \leq \ell \leq N-1, \tag{3.11}$$

and

$$\frac{\partial u}{\partial \nu}(y, \tilde{\rho}_\varepsilon(y)) = \frac{\varepsilon}{2} \frac{w'}{|y|} (\rho_\varepsilon)_{ij} y_i y_j + \frac{\varepsilon^2}{3} \frac{w'}{|y|} (\rho_\varepsilon)_{ijk} y_i y_j y_k + o(\varepsilon^2). \tag{3.12}$$

Using evenness-oddness properties of these functions, we see that

$$\int_{\partial R^N_+} w_\ell(y, 0) \frac{w'}{|y|} \rho_{ij} y_i y_j = 0,$$

and then, computing the integral on $\partial\Omega_\varepsilon$, we see that for any $R > 0$ we have

$$\int_{\partial\Omega_\varepsilon \cap B(0,R)} w_i(y, 0) \frac{w'}{|y|} (\rho_\varepsilon)_{ij} y_i y_j = O(\varepsilon^2).$$

We also see that

$$\int_{\partial R^N_+} w_\ell(y, 0) \frac{w'}{|y|} \rho_{ijk} y_i y_j y_k = K \rho_{ii\ell}, \tag{3.13}$$

where K is a nonzero constant. Then we conclude that

$$\frac{1}{\varepsilon^2} I_1 = \rho_{ii\ell} \hat{T}^\ell_0 + o(1). \tag{3.14}$$

From here and (3.10), taking the limit as $\varepsilon \to 0$ we find that

$$\nabla H(\bar{x}) \cdot \hat{T}_0 = 0 \tag{3.15}$$

and this contradicts hypothesis (H2). □

Finally we prove Theorem 1.1.

Proof of Theorem 1.1. By (3.1), we have that

$$u_\varepsilon(x) < a \quad \forall\ x \in \partial\Omega_0.$$

Hence u_ε satisfies (1.3) since $f(u_\varepsilon) = \bar{f}(u_\varepsilon)$ for $x \notin \Omega_0$. Since $\varepsilon^{-N} J_\varepsilon(u_\varepsilon) \to \frac{1}{2} I(w)$, by [19] or [23], we have that u_ε has only one local maximum point x_ε and $x_\varepsilon \in \Lambda$. By Lemma 3.2, $\lim_{\varepsilon \to 0} H(x_\varepsilon) = c$. The rest of the proof follows from [19] and [20]. □

Acknowledgments. The authors would like to thank the referees for carefully reading the manuscript. They pointed out several misprints and suggested some changes that made the text clearer. In particular, we thank them for the suggestion that simplified our hypotheses (H1)–(H2).

REFERENCES

[1] Adimurthi, G. Mancini, and S.L. Yadava, *The role of mean curvature in semilinear Neumann problem involving critical exponent*, Comm. Partial Differential Equations, 20 (1995), pp. 591–631.

[2] Adimurthi, F. Pacella, and S.L. Yadava, *Interaction between the geometry of the boundary and positive solutions of a semilinear Neumann problem with critical nonlinearity*, J. Funct. Anal., 113 (1993), pp. 8–350.

[3] Adimurthi, F. Pacella, and S.L. Yadava, *Characterization of concentration points and L^∞-estimates for solutions of a semilinear Neumann problem involving the critical Sobolev exponent*, Differential Integral Equations, 8 (1995), pp. 41–68.

[4] C.-C. Chen and C.-S. Lin, *Uniqueness of the ground state solution of $\Delta u + f(u) = 0$*, Comm. Partial Differential Equations, 16 (1991), pp. 1549–1572.

[5] M. del Pino and P.L. Felmer, *Semi-classical states for nonlinear Schrödinger equations*, J. Funct. Anal., 149 (1997), pp. 245–265.

[6] M. del Pino and P.L. Felmer, *Local mountain passes for semilinear elliptic problem in unbounded domains*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 121–137.

[7] M. del Pino and P. Felmer, *Multi-peak bound states for nonlinear Schrödinger equations*, Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 127–149.

[8] M. del Pino, P. Felmer, and J. Wei, *Multiple-peak solutions for some singular perturbation problems*, Calc. Var. Partial Differential Equations, to appear.

[9] M. del Pino, P. Felmer, and J. Wei, *On the role of the distance function in some singular perturbation problems*, Comm. Partial Differential Equations, to appear.

[10] C. Gui, *Multipeak solutions for a semilinear Neumann problem*, Duke Math. J., 84 (1996), pp. 739–769.

[11] C. Gui and N. Ghoussoub, *Multi-peak solutions for a semilinear Neumann problem involving the critical Sobolev exponent*, Math. Z., 229 (1998), pp. 443–474.

[12] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik (Berlin), 12 (1972), pp. 30–39.

[13] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in R^n*, in Mathematical Analysis and Applications, Part A, Adv. in Math. Suppl. Stud. 7A, Academic Press, New York, 1981, pp. 369–402.

[14] E.F. Keller and L.A. Segal, *Initiation of slime mold aggregation viewed as an instability*, J. Theory. Biol., 26 (1970), pp. 399–415.

[15] 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.

[16] Y.Y. Li, *On a singularly perturbed equation with Neumann boundary condition*, Comm. Partial Differential Equations, 23 (1998), pp. 487–545.

[17] C.-S. Liu, W.-M. Ni, and I. Takagi, *Large amplitude stationary solutions to a chemotaxis systems*, J. Differential Equations, 72 (1988), pp. 1–27.

[18] W.-M. Ni, X. Pan, and I. Takagi, *Singular behavior of least-energy solutions of a semilinear Neumann problem involving critical Sobolev exponents*, Duke Math. J., 67 (1992), pp. 1–20.

[19] W.-M. Ni and I. Takagi, *On the shape of least-energy solution to a semilinear Neumann problem*, Comm. Pure Appl. Math., 41 (1991), pp. 819–851.

[20] 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.

[21] W.-M. Ni and I. Takagi, *Point condensation generated by a reaction-diffusion system in axially symmetric domains*, Japan J. Industrial Appl. Math., 12 (1995), pp. 327–365.

[22] Z.-Q. Wang, *On the existence of multiple, single-peaked solutions for a semilinear Neumann problem*, Arch. Rational Mech. Anal., 120 (1992), pp. 375–399.

[23] J. Wei, *On the boundary spike layer solutions of a singularly perturbed semilinear Neumann problem*, J. Differential Equations, 134 (1997), pp. 104–133.

SIAM J. MATH. ANAL.
Vol. 31, No. 1, pp. 80–118

EXISTENCE OF NONPLANAR SOLUTIONS OF A SIMPLE MODEL OF PREMIXED BUNSEN FLAMES*

ALEXIS BONNET† AND FRANÇOIS HAMEL‡

Abstract. This work deals with the existence of solutions of a reaction-diffusion equation in the plane $\mathbb{R}^2$. The problem, whose unknowns are the real c and the function u, is the following:

$$(P) \qquad \begin{cases} \Delta u - c\dfrac{\partial u}{\partial y} + f(u) = 0 & \text{in } \mathbb{R}^2, \\ \forall \vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha), & u(\lambda\vec{k}) \underset{\lambda\to+\infty}{\longrightarrow} 0, \\ \forall \vec{k} \in \mathcal{C}(\vec{e}_2, \pi-\alpha), & u(\lambda\vec{k}) \underset{\lambda\to+\infty}{\longrightarrow} 1, \end{cases}$$

where $0 < \alpha \leq \pi/2$ is given, $\vec{e}_2 = (-1, 0)$, and, for any angle ϕ and any unit vector $\vec{e}$, $\mathcal{C}(\vec{e}, \phi)$ denotes the open half-cone with angle ϕ around the vector $\vec{e}$. The given function f is of the "ignition temperature" type. In this paper, we show the existence of a solution (c, u) of (P) and we give an explicit formula that relates the speed c and the angle α.

Key words. reaction-diffusion equations, sliding method, maximum principle, travelling waves, Bunsen flames

AMS subject classifications. 35B40, 35B50, 35J60, 35J65, 35Q35

PII. S0036141097316391

1. Introduction.

Bunsen flames are usually made of two flames: a diffusion flame and a premixed flame (see Figure 1 and the papers by Buckmaster and Ludford [11], Joulin [23], Liñan [27], and Sivashinsky [31], [32]). In this paper, we focus on the study of the premixed Bunsen flame. Roughly speaking, the hot products of the chemical reactions are located above the flame and the fresh gaseous mixture (fuel and oxidant) is located below (see Figure 1). For the sake of simplicity, we can assume that a global chemical reaction takes place in the gaseous mixture:

$$\mathcal{R}: \qquad \text{Fuel} \quad + O_2 \quad \rightarrow \quad \text{Products.}$$

The isotherms (level sets of the temperature) of the premixed Bunsen flame are conical in shape and, far away from the axis of symmetry, the flame is almost planar. The underlying subsonic mass flow goes upward from the fresh zone to the burnt gases with a uniform vertical velocity c.

In this paper, we deal with the stationary states of premixed flames that are invariant by translation in one of the directions orthogonal to the flow. Consequently, the mathematical problem only involves two variables (x, y) (see Figure 1). This situation occurs with Bunsen burners that have a thin rectangular cross section.

Under some additional physical conditions that correspond to the classical thermodiffusive model (see Berestycki and Larrouturou [4], Buckmaster and Ludford [11], Matkowsky and Sivashinsky [29]), the temperature $u(x, y)$, normalized in such a way

*Received by the editors February 12, 1997; accepted for publication (in revised form) November 19, 1998; published electronically November 10, 1999.

http://www.siam.org/journals/sima/31-1/31639.html

†Université de Cergy-Pontoise, 133 Boulevard du Port, 95011 Cergy-Pontoise Cedex, France (alexis.bonnet@mines.org).

‡Laboratoire d'Analyse Numérique, Tour 55-65, Université Paris VI, 4 place Jussieu, 75252 Paris Cedex 05, France (hamel@ann.jussieu.fr).

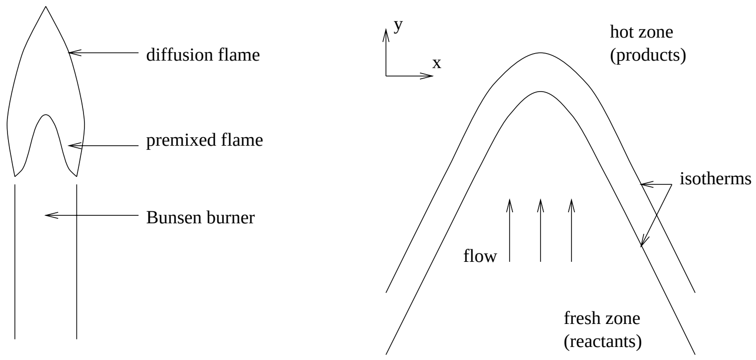


Fig. 1. *Bunsen flames (left) and the premixed flame (right).*

that $u \simeq 0$ in the fresh zone and $u \simeq 1$ in the hot zone far from the reaction sheet, solves the following reaction-diffusion equation in $\mathbb{R}^2 = \{(x,y),\ x \in \mathbb{R},\ y \in \mathbb{R}\}$:

$$\Delta u - c\frac{\partial u}{\partial y} + f(u) = 0 \ \text{ in } \mathbb{R}^2, \tag{1.1}$$

with the following limiting conditions at infinity:

$$\forall \vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha), \qquad u(\lambda\vec{k}) \underset{\lambda \to +\infty}{\longrightarrow} 0, \tag{1.2}$$

$$\forall \vec{k} \in \mathcal{C}(\vec{e}_2, \pi - \alpha), \qquad u(\lambda\vec{k}) \underset{\lambda \to +\infty}{\longrightarrow} 1, \tag{1.3}$$

where α is a given angle such that $0 < \alpha \leq \pi/2$. The vector $\vec{e}_2 = (0,1)$ is the unit vector in the direction $[Oy)$ and, for any unit vector $\vec{e}$ and any angle $\phi \in (0,\pi)$, $\mathcal{C}(\vec{e},\phi)$ denotes the open half-cone with aperture ϕ in the direction $\vec{e}$: $\mathcal{C}(\vec{e},\phi) = \{\vec{k} \in \mathbb{R}^2,\ \vec{k}\cdot\vec{e} > \|\vec{k}\|\ \|\vec{e}\| \cos\phi\}$. We also set $\mathcal{C}(z,\vec{e},\phi) = z + \mathcal{C}(\vec{e},\phi)$ for any point $z = (x,y) \in \mathbb{R}^2$.

The unknowns of this problem (1.1)–(1.3) are both the real c, which is like a nonlinear eigenvalue, and the function u, $0 < u < 1$, of class C^2 in $\mathbb{R}^2$. We shed light here on the fact that looking for the speed c, the angle α being known, is equivalent to looking for the angle α, the speed c being known, as is the case in experiments (see the comments after Theorem 1.2 below).

The function $1 - u$ also represents the relative concentration of the reactant. In (1.1), the terms Δu, $c\frac{\partial u}{\partial y}$, and $f(u)$ are, respectively, the diffusion, transport, and source terms. The source term $f(u)$, which may take into account the Arrhénius law and the mass action law, is given and Lipschitz continuous in $[0,1]$. Furthermore, one assumes that it is of the "ignition temperature" type:

$$\text{(1.4)}\ \exists\theta \in (0,1) \text{ such that } f \equiv 0 \text{ on } [0,\theta] \cup \{1\}, \quad f > 0 \text{ on } (\theta,1) \text{ and } f'(1) < 0.$$

For mathematical convenience, we extend f by 0 outside the interval $[0,1]$. The temperature θ is an ignition temperature, below which no chemical reaction happens.

In the one-dimensional case, the problem is reduced to

$$\begin{cases} u'' - c_0 u' + f(u) = 0, \\ u(-\infty) = 0,\ u(+\infty) = 1. \end{cases} \tag{1.5}$$

There have been many works devoted to the solutions of (1.5). We refer to the pioneering articles of Kolmogorov, Petrovsky, and Piskunov [26] for biological models, Zeldovich and Frank-Kamenetskii [37] for planar flames, as well as other papers by Aronson and Weinberger [2], Fife [14], Fife and McLeod [15], and Kanel' [24]. The main result is the following: if the function f fulfils (1.4), then there exist a unique real c_0 and a unique function $U(\xi)$ (up to translation with respect to ξ) which are solutions of (1.5). The real c_0 is positive and the function U is increasing in ξ. We may suppose that $U(0) = \theta$.

In more recent papers, multidimensional curved flames in infinite cylinders $\Sigma = \mathbb{R}\times\omega = \{(x_1, y),\, x_1 \in \mathbb{R},\, y \in \omega\}$, with smooth cross sections ω, have been investigated. In this case, the temperature $u(x, y)$ solves the equations

$$\left\{\begin{array}{rll} \Delta u - (c + \alpha(y))\dfrac{\partial u}{\partial x_1} + f(u) & = 0 & \text{in } \Sigma, \\ u(-\infty, \cdot) = 0,\ u(+\infty, \cdot) & = 1, & \\ \dfrac{\partial u}{\partial \nu} & = 0 & \text{on } \partial\Sigma, \end{array}\right. \tag{1.6}$$

where ν is the outward unit normal to $\partial\omega$ and $\alpha(y)$ is the x_1-component of the given underlying flow (see Berestycki and Larrouturou [5]; Berestycki, Larrouturou, and Lions [6]; Berestycki and Nirenberg [9]; Vega [33]; Volpert and Volpert [34]; and Xin [36] under periodic conditions). If $\alpha(y) = \alpha_0$ does not depend on y, it is known that (1.6) has a unique solution and that it is planar; namely, it depends only on the longitudinal variable x_1. If the function $y \mapsto \alpha(y)$ is not constant, the solution of (1.6) still exists and is unique, but it is not planar anymore (such solutions correspond to curved flames). Nonplanar flames may also be observed in infinite cylinders under different physical conditions: Glangetas and Roquejoffre [18] and Margolis and Sivashinsky [28] proved that if the single partial differential equation in (1.6) was replaced with a system of two reaction-diffusion equations, then a bifurcation toward nonplanar flames might occur.

Let us now come back to the question of the existence of solutions (c, u) of the problem (1.1)–(1.3). If $\alpha = \pi/2$, the couple (c_0, U) is obviously a solution. The question of the existence of solutions if $\alpha < \pi/2$ has so far remained open. In this paper, we show the existence of a speed c and of a nonplanar—if $\alpha < \pi/2$—function u defined in $\mathbb{R}^2$, which are solutions of (1.1)–(1.3). As a consequence, nonplanar flames exist for the model (1.1)–(1.3) although this model involves only one reaction-diffusion equation (and not two such equations) and although the underlying flow is uniform.

In this paper, we prove two main theorems. The first one states the existence of a solution (c, u) of (1.1)–(1.3) for any angle $0 < \alpha \le \pi/2$. The second one deals with the question of the speed c's uniqueness.

Theorem 1.1. *Let f fulfill* (1.4) *("ignition temperature" profile). For any $\alpha \in (0, \pi/2]$, there exists a solution (c, u) of* (1.1)–(1.3)*, namely,*

$$\left\{\begin{array}{lrl} & \Delta u - c\dfrac{\partial u}{\partial y} + f(u) = & 0 \quad \textit{in } \mathbb{R}^2, \\ \forall \vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha), & u(\lambda\vec{k}) \underset{\lambda\to+\infty}{\longrightarrow} & 0, \\ \forall \vec{k} \in \mathcal{C}(\vec{e}_2, \pi - \alpha), & u(\lambda\vec{k}) \underset{\lambda\to+\infty}{\longrightarrow} & 1, \end{array}\right.$$

such that

$$c = \frac{c_0}{\sin\alpha}. \tag{1.7}$$

Furthermore, $0 < u < 1$, u *is symmetric with respect to the variable* x, *and* u *is decreasing in any direction* $\vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha)$. *The following limiting conditions, which are stronger than* (1.2)–(1.3), *also hold:*

$$u(\lambda \vec{k}') \to 0 \quad \text{as } \lambda \to +\infty \text{ and } \vec{k}' \to \vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha), \tag{1.8}$$

$$u(\lambda \vec{k}') \to 1 \quad \text{as } \lambda \to +\infty \text{ and } \vec{k}' \to \vec{k} \in \mathcal{C}(\vec{e}_2, \pi - \alpha). \tag{1.9}$$

Finally, for each $\lambda \in (0,1)$, *the level set* $\{(x,y),\ u(x,y) = \lambda\}$ *is a curve* $\{y = \varphi_\lambda(x), x \in \mathbb{R}\}$ *and it has two asymptotic directions that are directed by the vectors* $(\pm \sin\alpha, -\cos\alpha)$. *If* $x_n \to -\infty$, *then the functions* $u_n(x,y) = u(x + x_n, y + \varphi_\lambda(x_n))$ *converge locally to the planar function* $U(y \sin\alpha - x\cos\alpha + U^{-1}(\lambda))$.

THEOREM 1.2. *Let* f *fulfill* (1.4) *and* α *be an angle in* $(0, \pi/2]$. *If* (c,u) *is a solution of* (1.1) *and* (1.8)–(1.9), *then*

$$c = \frac{c_0}{\sin\alpha}.$$

We can see that the speed $c = c_0/\sin\alpha$ of the nonplanar flame (for $\alpha < \pi/2$) is greater than the speed c_0 of the planar flame. Furthermore, the angle α is all the smaller as the speed c is larger. That is physically meaningful since the curvature of the flame increases with the speed of the fuel flow. It is worth noticing that the formula (1.7) has been known for a long time and had been formally derived from the planar behavior of the flame, far away from its center, along the directions $(\pm \sin\alpha, -\cos\alpha)$. This formula had been used in experiments to find the planar speed c_0: indeed, the vertical speed c of the gases at the exit of the Bunsen burner being known, one can measure the angle α and the one-dimensional speed c_0 is then given by the formula $c_0 = c \sin\alpha$ (see [31], Williams [35]).

Hence, the results of Theorems 1.1 and 1.2 are not surprising. Nevertheless, they are the first rigorous analysis of the conical premixed Bunsen flames.

REMARK 1.3. *From Theorem* 1.1, *there is a continuum of solutions* $(c_0/\sin\alpha, u)$ *solving* (1.1) *and satisfying the simple asymptotic limits* $u(x, -\infty) = 0$ *and* $u(x, +\infty) = 1$ *for all* $x \in \mathbb{R}$. *This is in contrast with problem* (1.6) *mentioned above. However, if the limits* $u(x,-\infty) = 0$ *and* $u(x, +\infty) = 1$ *are uniform with respect to* $x \in \mathbb{R}$, *then* (c_0, U) *will be the unique solution of* (1.1) *up to translation in the variables* (x, y) *for* U *(see Hamel and Monneau* [21]*).*

Open questions.

(1) For each fixed angle $\alpha \in]0, \pi/2]$, do all the solutions u of (1.1)–(1.3) have the same profile? What kind of a priori monotonicity or symmetry properties do they fulfill? Are they stable for the evolution problem $\partial_t u = \Delta u - c\partial_y u + f(u)$? Answers to some of those questions are given in [21].

(2) Is there any solution (c,u) to (1.1)–(1.3) if $\alpha > \pi/2$? The answer is no and is given in [21].

(3) Is there any solution (c,u) to the free boundary problem equivalent to (1.1)–(1.3) and obtained in the limit of "high activation energies"? The answer is yes (see Hamel and Monneau [22]).

(4) Are there three-dimensional flames and, if so, are they necessarily invariant by rotation?

Structure of the paper. Section 2 is devoted to solving problems that are similar to (1.1)–(1.3) but are set in finite rectangles $[-a, a] \times [-a \cot\gamma_a, a \cot\gamma_a]$ where

γ_a is an angle close to α. For those problems, some a priori estimates about the speeds c_a and the functions u_a are established. A technical lemma, which is proved in the Appendix (section 5), is devoted to determining the behavior of the functions u_a near the corners of the rectangles. In section 3, we pass to the limit $a \to \infty$ in the whole plane and we determine the shape of the level sets of the limit function u by resorting to arguments of the "sliding method" type. In section 4, we prove Theorem 1.2.

REMARK 1.4. *The proof of Theorem* 1.1*, which is detailed in the next sections, actually allows us to get an independent result about the following problem set in an infinite strip* $\Sigma = \{(x,y) \in (-L,L) \times \mathbb{R}\}$ *with oblique Neumann boundary conditions:*

$$\begin{cases} \Delta u - c\partial_y u + f(u) = & 0 \text{ in } \Sigma, \\ \forall y \in \mathbb{R}, \quad \partial_\tau u(-L,y) = \partial_{\tilde{\tau}} u(L,y) = & 0, \\ u(\cdot,-\infty) = 0, \ u(\cdot,+\infty) = & 1, \end{cases} \tag{1.10}$$

where $\tau = (-\sin\alpha, -\cos\alpha)$ *and* $\tilde{\tau} = (\sin\alpha, -\cos\alpha)$*. Namely, with the same method as for Theorem* 1.1*, it follows that there exists a solution* (c,u) *to* (1.10) *such that the function* u *is nondecreasing in each direction* $\rho \in \overline{\mathcal{C}(\vec{e}_2,\alpha)}$*.*

2. Solving equivalent problems in finite rectangles. Let us set any real $a > 1/\alpha^2$ and $\gamma_a = \alpha - 1/\sqrt{a}$. The angle γ_a is such that $0 < \gamma_a < \alpha$, $\gamma_a \to \alpha$ and $a(\cot\gamma_a - \cot\alpha) \to +\infty$ as $a \to +\infty$. Let Σ_a be the bounded and open rectangle $\Sigma_a = (-a,a)\times(-a\cot\gamma_a, a\cot\gamma_a)$. Call $\tau = (-\sin\alpha, -\cos\alpha)$ and $\tilde{\tau} = (\sin\alpha, -\cos\alpha)$ (see Figure 2). When there is no confusion, γ_a is often replaced with γ.

In this section, we focus on the questions of the existence and the uniqueness as well as on a priori estimates of the solutions (c_a, u_a) to the following problem:

$$\begin{cases} \Delta u_a - c_a\partial_y u_a + f(u_a) = & 0 \text{ in } \Sigma_a, \\ \forall x \in [-a,a], \quad u_a(x,-a\cot\gamma_a) = 0, \ u_a(x,a\cot\gamma_a) = & 1, \\ \forall y \in (-a\cot\gamma_a, a\cot\gamma_a), \quad \dfrac{\partial u_a}{\partial\tau}(-a,y) = \dfrac{\partial u_a}{\partial\tilde{\tau}}(a,y) = & 0 \end{cases} \tag{2.1}$$

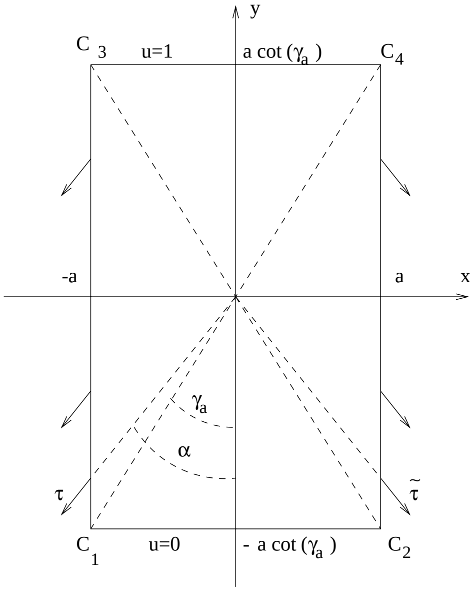


FIG. 2. *The rectangle* Σ_a*.*

under the following normalization condition:

$$\max_{\substack{y=-\cot\alpha\ \ |x| \\ -a\le x\le a}} u_a(x,y)=\theta. \tag{2.2}$$

2.1. Existence of solutions of (2.1)–(2.2) and a priori bounds for the speeds c_a.

2.1.1. On the solutions u_c of (2.1). Let c be any fixed real. Let us call $(C_i)_{1\le i\le 4}$ the four corners of Σ_a: $C_1=(-a,-a\cot\gamma)$, $C_2=(a,-a\cot\gamma)$, $C_3=(-a,a\cot\gamma)$, $C_4=(a,a\cot\gamma)$ (see Figure 2) and set $\tilde{\Sigma}_a=\overline{\Sigma_a}\setminus\cup_{i=1}^{4}\{C_i\}$.

Now consider the following Dirichlet–Neumann problem:

$$\left\{\begin{array}{r}\Delta u-c\partial_y u+f(u)=0\qquad\text{in }\Sigma_a,\\ \forall x\in[-a,a],\quad u(x,-a\cot\gamma)=0,\ u(x,a\cot\gamma)=1,\\ \forall y\in(-a\cot\gamma,a\cot\gamma),\quad \partial_\tau u(-a,y)=\partial_{\tilde{\tau}}u(a,y)=0.\end{array}\right. \tag{2.3}$$

This problem is the same as (2.1), but the speed c is given in (2.3) and only the function u is unknown. The following three lemmas are similar to some of the results in a paper by Berestycki and Nirenberg [7]. The proofs, which will be used several times in the sequel, are written for the sake of completeness.

LEMMA 2.1. *For each speed $c\in\mathbb{R}$, we have that problem* (2.3) *has a solution u_c in $\cap_{p>1}W^{2,p}_{loc}(\tilde{\Sigma}_a)\cap\ C(\overline{\Sigma_a})$, where $C(\overline{\Sigma_a})$ is the space of all continuous functions in $\overline{\Sigma_a}$.*

Proof. Let $(\Sigma_{a,\varepsilon})_{\varepsilon>0}$ be a sequence of bounded and smooth domains such that, for each $\varepsilon>0$,

$$\Sigma_a\setminus\bigcup_{i=1}^{4}B(C_i,\varepsilon)\ \subset\ \Sigma_{a,\varepsilon}\ \subset\ \Sigma_a,$$

where $B(C_i,\varepsilon)$ denotes the open ball centered on the point C_i with radius ε. Let $\varepsilon>0$ be small enough. Consider a smooth vector field $\rho_\varepsilon(x,y)$ defined on $\partial\Sigma_{a,\varepsilon}$ such that $\rho_\varepsilon\cdot\nu_\varepsilon\ge 0$ on $\partial\Sigma_{a,\varepsilon}$ (where ν_ε is the outward unit normal to $\partial\Sigma_{a,\varepsilon}$) $\rho_\varepsilon=\tau$ on $\{-a\}\times(-a\cot\gamma\ +\varepsilon,a\cot\gamma\ -\varepsilon)$, $\rho_\varepsilon=\tilde{\tau}$ on $\{a\}\times(-a\cot\gamma\ +\varepsilon,a\cot\gamma\ -\varepsilon)$, and $\rho_\varepsilon=\vec{0}$ on $(-a+\varepsilon,a-\varepsilon)\times\{\pm a\cot\gamma\}$. Let $\sigma_{0,\varepsilon}(x,y)$ be a smooth nonnegative function defined on $\partial\Sigma_{a,\varepsilon}$ such that $\sigma_{0,\varepsilon}=1$ on $\partial\Sigma_{a,\varepsilon}\cap\{y\le -a\cot\gamma\ +\varepsilon\}$ and $\sigma_{0,\varepsilon}=0$ on $\partial\Sigma_{a,\varepsilon}\cap\{y\ge -a\cot\gamma\ +2\varepsilon\}$. Last, let $\sigma_{1,\varepsilon}$ be a smooth nonnegative function defined on $\partial\Sigma_{a,\varepsilon}$ such that $\sigma_{1,\varepsilon}=1$ on $\partial\Sigma_{a,\varepsilon}\cap\{y\ge a\cot\gamma\ -\varepsilon\}$ and $\sigma_{1,\varepsilon}=0$ on $\partial\Sigma_{a,\varepsilon}\cap\{y\le a\cot\gamma\ -2\varepsilon\}$. For each $\varepsilon>0$ small enough, the problem

$$\left\{\begin{array}{r}\Delta u_\varepsilon-c\partial_y u_\varepsilon+f(u_\varepsilon)=0\quad\text{in }\Sigma_{a,\varepsilon},\\ \rho_\varepsilon\cdot\nabla u+\sigma_{0,\varepsilon}u+\sigma_{1,\varepsilon}(u-1)=0\quad\text{on }\partial\Sigma_{a,\varepsilon}\end{array}\right.$$

has a solution u_ε such that $0\le u_\varepsilon\le 1$ since 0 and 1, respectively, are sub- and supersolutions (see Berestycki and Nirenberg [7]).

From the standard elliptic estimates up to the boundary (Agmon, Douglis, and Nirenberg [1]; Gilbarg and Trudinger [17]), up to extraction of some subsequence, the functions u_ε approach a function $u_c\in\bigcap\limits_{p>1}W^{2,p}_{loc}(\tilde{\Sigma}_a)\ \cap\ C_{loc}(\tilde{\Sigma}_a)$ as $\varepsilon\to 0$. The function u_c is a solution of

$$\left\{\begin{array}{r}\Delta u_c-c\partial_y u_c+f(u_c)=0\qquad\text{in }\Sigma_a,\\ \forall x\in(-a,a),\quad u_c(x,-a\cot\gamma)=0,\ u_c(x,a\cot\gamma)=1,\\ \forall y\in(-a\cot\gamma,a\cot\gamma),\quad \partial_\tau u_c(-a,y)=\partial_{\tilde{\tau}}u_c(a,y)=0.\end{array}\right. \tag{2.4}$$

Furthermore, we claim that, for each $i \in \{1, \dots, 4\}$, there exists a function $\overline{v}_i$ defined in a neighborhood V_i of the corner C_i such that $\overline{v}_i(C_i) = 0$ and, for all $\varepsilon > 0$ small enough,

$$\begin{array}{lrl} \text{if } i = 1 \text{ or } 2, & u_\varepsilon(x,y) & \leq \overline{v}_i(x,y) \\ \text{if } i = 3 \text{ or } 4, & 1 - u_\varepsilon(x,y) & \leq \overline{v}_i(x,y) \end{array} \quad \text{in } V_i \cap \overline{\Sigma_{a,\varepsilon}}. \tag{2.5}$$

The proof of this fact is temporarily postponed and will be given in Remark 5.2 in section 5.

As a consequence, the function u_c can be extended by continuity at the four corners C_i of Σ_a. In other words, $u_c \in \cap_{p>1} W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C(\overline{\Sigma_a})$. From the strong maximum principle and the Hopf lemma, it also follows that $0 < u_c < 1$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$. □

Lemma 2.2. *The function u_c is increasing in y and it is the unique solution of (2.3) in $\cap_{p>1} W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C(\overline{\Sigma_a})$. Furthermore, if f is of class C^1, then $\partial_y u_c > 0$ in $\tilde{\Sigma}_a$.*

Proof. It is based on the sliding method (see [7]). Let u be any solution of (2.3) in $\cap_{p>1} W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C(\overline{\Sigma_a})$. For any $\lambda \in (0, 2a\cot\gamma)$, let v^λ be the function defined by $v^\lambda(x,y) = u(x, y-\lambda) - u(x,y)$ in the set

$$\Sigma^\lambda_a = (-a,a) \times (-a\cot\gamma + \lambda, a\cot\gamma). \tag{2.6}$$

Since u is uniformly continuous on the compact set $\overline{\Sigma_a}$ and since $u(\cdot, -a\cot\gamma) = 0$, $u(\cdot, a\cot\gamma) = 1$, there exists $\varepsilon > 0$ small enough such that v^λ is negative in $\overline{\Sigma^\lambda_a}$ for all λ in $[2a\cot\gamma - \varepsilon, 2a\cot\gamma)$.

Let us now decrease λ. Suppose that there exists $\lambda^* > 0$ such that $v^\lambda < 0$ in $\overline{\Sigma^\lambda_a}$ for all $\lambda \in (\lambda^*, 2a\cot\gamma)$ and $v^{\lambda^*} \leq 0$ in $\overline{\Sigma^{\lambda^*}_a}$ with equality somewhere at a point $(\overline{x}, \overline{y}) \in \overline{\Sigma^{\lambda^*}_a}$. Since $0 < u < 1$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$, the function v^{λ^*} is negative at the "bottom" $[-a,a] \times \{-a\cot\gamma + \lambda^*\}$ of the boundary of $\Sigma^{\lambda^*}_a$. Similarly, the function v^{λ^*} is negative at the "top" $[-a,a] \times \{a\cot\gamma\}$ of the boundary of $\Sigma^{\lambda^*}_a$. We also have $\partial_\tau v^{\lambda^*}(-a,y) = \partial_{\tilde{\tau}} v^{\lambda^*}(a,y) = 0$ for all $y \in (-a\cot\gamma + \lambda^*, a\cot\gamma)$. The nonpositive function v^{λ^*} satisfies the elliptic equation

$$\Delta v^{\lambda^*} - c\partial_y v^{\lambda^*} + c(x,y) v^{\lambda^*} = 0 \quad \text{in } \Sigma^{\lambda^*}_a,$$

where the function $c(x,y)$ is bounded in $\Sigma^{\lambda^*}_a$ because of the Lipschitz continuity of f. Since $v^{\lambda^*}(\overline{x}, \overline{y}) = 0$ at a point $(\overline{x}, \overline{y}) \in \overline{\Sigma^{\lambda^*}_a}$, we then conclude from the strong maximum principle (if $-a < \overline{x} < a$) or from the Hopf lemma (if $\overline{x} = \pm a$) that $v^{\lambda^*} \equiv 0$ in $\overline{\Sigma^{\lambda^*}_a}$. That is ruled out by the boundary conditions on $[-a,a] \times \{-a\cot\gamma + \lambda^*, a\cot\gamma\}$.

Hence, there is no such $\lambda^* > 0$. We finally conclude that

$$\forall 0 < \lambda < 2a\cot\gamma, \quad u^\lambda(x,y) = u(x, y-\lambda) < u(x,y) \ \text{ in } \overline{\Sigma^\lambda_a}.$$

This yields that for any $x \in [-a,a]$, the function $y \mapsto u(x,y)$ is strictly increasing with respect to $y \in [-a\cot\gamma, a\cot\gamma]$.

If f is of class C^1, we can differentiate the equation satisfied by u. From the strong maximum principle and the Hopf lemma, it follows that $\partial_y u > 0$ in $\tilde{\Sigma}_a$.

The second part of Lemma 2.2, namely, the uniqueness of the solution u_c of (2.3) in $\cap_{p>1} W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C(\overline{\Sigma_a})$, could be proved in the same way. Indeed, if there were two solutions u_c and u'_c, we would find as above that $u_c(x, y-\lambda) < u'_c(x,y)$ in $\overline{\Sigma^\lambda_a}$ for all $\lambda \in (0, 2a\cot\gamma)$, whence $u_c \le u'_c$ in $\overline{\Sigma_a}$. Changing u_c and u'_c, we have $u'_c \le u_c$ and finally $u_c = u'_c$. □

Corollary 2.3. *For each c, the function u_c is symmetric with respect to x.*

Proof. Indeed, if u_c denotes the unique solution of (2.3), the function $\tilde{u}(x,y) = u_c(-x,y)$ is also a solution. By uniqueness, we have $\tilde{u} = u_c$. □

Lemma 2.4. *The functions u_c are decreasing and continuous, with respect to c, in the spaces $W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C(\overline{\Sigma_a})$ in the following sense: if $c < c'$, then $u_c > u_{c'}$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$ and if $c \to c_0$, then $u_c \to u_{c_0}$ in $\cap_{p>1} W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C(\overline{\Sigma_a})$.*

Proof. Choose any c and c' such that $c < c'$. We have to prove that $u_c > u_{c'}$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$. For each $0 < \lambda < 2a\cot\gamma$, we define the function $v^\lambda(x,y) = u_{c'}(x, y-\lambda) - u_c(x,y)$ in Σ_a^λ (see definition (2.6)).

If λ is close enough to $2a\cot\gamma$, we have $v^\lambda < 0$ in $\overline{\Sigma_a^\lambda}$ thanks to the boundary conditions fulfilled by u_c and $u_{c'}$. Let us now suppose that there exists $\lambda^* > 0$ such that $v^\lambda < 0$ in $\overline{\Sigma_a^\lambda}$ for all $\lambda \in (\lambda^*, 2a\cot\gamma)$ and $v^{\lambda^*} \le 0$ with equality somewhere in $\overline{\Sigma_a^{\lambda^*}}$. The function v^{λ^*} satisfies

$$(2.7) \qquad \begin{cases} \Delta v^{\lambda^*} - c\partial_y v^{\lambda^*} + c(x,y)v^{\lambda^*} & = (c'-c)\partial_y u_{c'}(x, y-\lambda^*) \ \text{ in } \Sigma_a^{\lambda^*}, \\ \quad \partial_\tau v^{\lambda^*}(-a,y) = \partial_{\tilde{\tau}} v^{\lambda^*} & = 0 \quad \forall y \in (-a\cot\gamma + \lambda^*, a\cot\gamma) \end{cases}$$

for a bounded function $c(x,y)$. On the one hand, since $c < c'$ and $\partial_y u_{c'} \ge 0$ (from the first part of Lemma 2.2), it follows from the strong maximum principle and the Hopf lemma that $v^{\lambda^*} \equiv 0$ in $\overline{\Sigma_a^{\lambda^*}}$. On the other hand, since $0 < u_c, u_{c'} < 1$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$, we have $v^{\lambda^*} < 0$ on $[-a,a] \times \{-a\cot\gamma + \lambda^*, a\cot\gamma\}$. That eventually leads to a contradiction.

Hence, for all $\lambda \in (0, 2a\cot\gamma)$, we have

$$v^\lambda = u_{c'}(x, y-\lambda) - u_c(x,y) < 0 \text{ in } \overline{\Sigma_a^\lambda}.$$

Then, $u_c \ge u_{c'}$ in $\overline{\Sigma_a}$. Since $v^0 = u_{c'} - u_c$ satisfies equation (2.7), the strong maximum principle and the Hopf lemma yield that $u_c > u_{c'}$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$.

Now, consider a sequence (c_n) such that $c_n \to c_0 \in \mathbb{R}$ as $n \to +\infty$. From the standard elliptic estimates up to the boundary, and up to extraction of some subsequence, the functions u_{c_n} approach a function $\tilde{u}_{c_0} \in \cap_{p>1} W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C_{loc}(\tilde{\Sigma}_a)$. The function $\tilde{u}_{c_0}$ is a solution of (2.4) with the speed c_0. Furthermore, for each $i \in \{1, \ldots, 4\}$, there exists a function $\overline{v}_i$ defined in a neighborhood V_i of the corner C_i, such that $\overline{v}_i(C_i) = 0$ and, for n large enough,

$$(2.8) \qquad \begin{array}{llll} \text{if } i = 1 \text{ or } 2, & u_{c_n}(x,y) & \le \overline{v}_i(x,y) & \multirow{2}{*}{} \\ \text{if } i = 3 \text{ or } 4, & 1 - u_{c_n}(x,y) & \le \overline{v}_i(x,y) & \end{array} \quad \text{in } V_i \cap \overline{\Sigma_a}$$

(see Remark 5.2). Hence, the function $\tilde{u}_{c_0}$ can be extended by continuity at the four corners C_i. As a consequence, $\tilde{u}_{c_0} = u_{c_0}$. Furthermore, since the functions u_{c_n} approach u_{c_0} in any compact subset of $\tilde{\Sigma}_a$, the above estimates around the four corners C_i also imply that u_{c_n} approach u_{c_0} uniformly in $\overline{\Sigma_a}$. Finally, since the limit function u_{c_0} is unique, it follows that the whole sequence (u_{c_n}) approaches u_{c_0} as $n \to +\infty$. □

2.1.2. Estimating the speeds.

In this subsection, we aim at establishing some a priori estimates for the speeds c_a of the possible solutions (c_a, u_a) of (2.1)–(2.2).

We first need some preliminary results about the speeds of some one-dimensional traveling fronts. Remember that the function f has been extended by 0 outside $[0,1]$.

Let $f'_-(1) = \lim_{t\to 1,\ t<1} \frac{f(t)}{t-1}$. For each $0 < \eta < \min(1-\theta, |f'_-(1)|)$, let f_η be a C^1 function in $[0,1]$, fulfilling (1.4) with the ignition temperature $\theta + \eta$, such that $f'_\eta(1) = f'_-(1) + \eta$, $f - \eta \le f_\eta \le f$ in $[0,1]$, and $f_\eta < f$ in $(\theta, 1)$. As for f, we also extend f_η by 0 outside $[0,1]$. From the results in [2], [9], [15] and [24], there exists a unique real c_0^η and a unique function u_η solving

$$\left\{\begin{array}{rl} u''_\eta - c_0^\eta u'_\eta + f_\eta(u_\eta) = 0 & \text{in } \mathbb{R}, \\ u_\eta(-\infty) = -\eta,\ u_\eta(0) = \theta,\ u_\eta(+\infty) = 1. & \end{array}\right.$$

Moreover, $u'_\eta > 0$ in $\mathbb{R}$. With the same arguments as in the paper by Berestycki and Nirenberg [9], it also follows that $c_0^\eta \overset{\le}{\to} c_0$ as $\eta \to 0$ (remember that c_0 is the unique speed for which (1.5) has a solution).

LEMMA 2.5. *Under the above notation, there exists a real $a_1(\eta) > 0$ such that if $a \ge a_1(\eta)$ and if $c < c_0^\eta/\sin\alpha$, then $\theta < \max_{\substack{y=-\cot\alpha\ |x| \\ |x|\le a}} u_c$.*

Proof. Assume that c is such that $c < c_0^\eta/\sin\alpha$. Let u_c be the solution of (2.3) and set $v(x,y) = u_\eta(\cos\alpha\ x + \sin\alpha\ y)$ in Σ_a. We want to prove that if a is large enough, then this function v is a subsolution of problem (2.3).

We have

$$\begin{array}{rl} \Delta v - c\partial_y v + f(v) & = u''_\eta - c\sin\alpha\ u'_\eta + f(u_\eta) \\ & = (c_0^\eta - c\sin\alpha)u'_\eta(\cos\alpha\ x + \sin\alpha\ y) + f(u_\eta) - f_\eta(u_\eta) \\ & > 0 \ \text{ in } \Sigma_a \end{array}$$

since $c < c_0^\eta/\sin\alpha$, $u'_\eta > 0$, and $f \ge f_\eta$. Furthermore, for all $y \in (-a\cot\gamma_a, a\cot\gamma_a)$, we can see that

$$\partial_\tau v(-a,y) = -2\sin\alpha\cos\alpha\ u'_\eta(-a\cos\alpha + \sin\alpha\ y) \le 0$$

and that $\partial_{\tilde\tau} v(a,y) = 0$. At the "top"of the boundary of Σ_a, we have $v(x, a\cot\gamma_a) < 1$ for all $x \in [-a,a]$. At the "bottom" of the boundary of Σ_a, the function v is equal to

$$v(x, -a\cot\gamma_a) = u_\eta(\cos\alpha\ x - a\cot\gamma_a\ \sin\alpha).$$

Since $|x| \le a$, it follows that

$$\cos\alpha\ x - a\cot\gamma_a\sin\alpha \le (\cos\alpha - \cot\gamma_a\sin\alpha)\ a \to -\infty \text{ as } a \to +\infty$$

since $\gamma_a = \alpha - 1/\sqrt{a}$ for $a > 1/\alpha^2$. On the other hand, the function u_η is increasing and $u_\eta(\xi) \to -\eta$ as $\xi \to -\infty$. Consequently, there exists a real $a_1(\eta)$ such that

$$(a \ge a_1(\eta)) \implies (\forall x \in [-a,a],\ v(x, -a\cot\gamma) < 0).$$

Hence, if $c < c_0^\eta/\sin\alpha$ and if $a \ge a_1(\eta)$, the function v is a subsolution of problem (2.3). Remember now that the function u_c is a solution of (2.3). As in the proof of the monotonicity result in Lemma 2.2, we can compare the functions v and u_c by using a sliding method. We would find that $v < u_c$ in Σ_a. This yields that $v(0,0) = \theta < u_c(0,0)$, whence $\theta < \max_{\substack{y=-\cot\alpha\ |x| \\ |x|\le a}} u_c$. That completes the proof of Lemma 2.5. □

The next lemma states that if the speed c is large enough, then the solution u_c of (2.3) will be below θ on the set $\{y = -\cot\alpha\ |x|,\ |x| \le a\}$. Before doing that, we need a few auxiliary notation. For any $\varepsilon \in (0,\theta)$, let f^ε be a C^1 function in

$[0,1+\varepsilon]$ such that $f^\varepsilon \equiv 0$ in $(-\infty,\theta-\varepsilon]\cup[1+\varepsilon,+\infty)$, $f^\varepsilon > 0$ in $(\theta-\varepsilon,1+\varepsilon)$, $(f^\varepsilon)'_-(1+\varepsilon) := \lim_{t\to 1+\varepsilon,\ t<1+\varepsilon} \frac{f^\varepsilon(t)}{t-1-\varepsilon}$ exists and is negative. In other words, f^ε fulfills the assertion (1.4) on the interval $[0,1+\varepsilon]$ with the ignition temperature $\theta-\varepsilon$. Moreover, one assumes that $f \le f^\varepsilon \le f+\varepsilon$ in $\mathbb{R}$ and $f < f^\varepsilon$ in $[\theta,1]$. From the results in [2], [9], [15] and [24], there exists a unique real $\tilde c_0^\varepsilon$ and a unique function u^ε defined in $\mathbb{R}$ such that

$$\left\{\begin{array}{ll} (u^\varepsilon)'' - \tilde c_0^\varepsilon (u^\varepsilon)' + f^\varepsilon(u^\varepsilon) & = 0 \text{ in } \mathbb{R}, \\ u^\varepsilon(-\infty)=0,\ u^\varepsilon(0)=\theta,\ u^\varepsilon(+\infty) & = 1+\varepsilon. \end{array}\right.$$

Moreover, one has $(u^\varepsilon)' > 0$ in $\mathbb{R}$ and $\tilde c_0^\varepsilon \overset{>}{\to} c_0$ as $\varepsilon \to 0$ (see [9]).

LEMMA 2.6. *There exists a real $a_2(\varepsilon)$ such that if $a \ge a_2(\varepsilon)$ and if $c > \tilde c_0^\varepsilon/\sin^2\alpha$, then $\theta > \max_{\substack{y=-\cot\alpha\ |x| \\ |x|\le a}} u_c$.*

Proof. Let c be a real such that $c > \tilde c_0^\varepsilon/\sin^2\alpha$. Let us set

$$\beta = \frac{3\cot\alpha}{2(c-\tilde c_0^\varepsilon/\sin^2\alpha)}$$

and choose $a > \beta$. Let us call φ the function defined in $\mathbb{R}$ by

$$\left\{\begin{array}{ll} \varphi(x) = \dfrac{\cot\alpha}{8\beta^3}x^4 - \dfrac{3\cot\alpha}{4\beta}x^2 & \text{if } |x|\le\beta, \\ \varphi(x) = -|x|\cot\alpha + \dfrac{3}{8}\beta\cot\alpha & \text{if } \beta\le|x|\le a. \end{array}\right.$$

It is easy to see that the function φ is concave, is of class C^2 in $\mathbb{R}$, and that $|\varphi'(x)| \le \cot\alpha$, $|\varphi''(x)| \le c - \tilde c_0^\varepsilon/\sin^2\alpha$.

Let us now define the function $v(x,y) = u^\varepsilon(y-\varphi(x))$ in Σ_a and check that this function v is a supersolution of (2.3) for a large enough. We have

$$\partial_y v = (u^\varepsilon)'(y-\varphi(x))$$

and $$\Delta v = (1+\varphi'(x)^2)(u^\varepsilon)''(y-\varphi(x)) - \varphi''(x)(u^\varepsilon)'(y-\varphi(x)).$$

Hence,

$$\begin{array}{ll} \Delta v - c\partial_y v + f(v) & = (1+\varphi'(x)^2)(u^\varepsilon)''(y-\varphi(x)) \\ & \quad -(c+\varphi''(x))(u^\varepsilon)'(y-\varphi(x)) + f(u^\varepsilon(y-\varphi(x))) \\ & = [\tilde c_0^\varepsilon(1+\varphi'(x)^2) - c - \varphi''(x)]\ (u^\varepsilon)'(y-\varphi(x)) \\ & \quad -\varphi'(x)^2 f^\varepsilon(u^\varepsilon(y-\varphi(x))) \\ & \quad +f(u^\varepsilon(y-\varphi(x))) - f^\varepsilon(u^\varepsilon(y-\varphi(x))). \end{array}$$

On the one hand, we know that $(u^\varepsilon)' > 0$ and that $0 \le f \le f^\varepsilon$. On the other hand, in view of the definition of φ, we infer that

$$\forall x \in \mathbb{R}, \quad \tilde c_0^\varepsilon(1+\varphi'(x)^2) - c - \varphi''(x) \le 0.$$

It follows that

$$\Delta v - c\partial_y v + f(v) \le 0 \ \text{ in } \Sigma_a.$$

Furthermore, one has, for all $y \in (-a\cot\gamma_a, a\cot\gamma_a)$,

$$\begin{array}{ll} \partial_\tau v(-a,y) & = (\sin\alpha\ \varphi'(-a) - \cos\alpha)\ (u^\varepsilon)'(y-\varphi(-a)) \\ & = 0 \end{array}$$

since $\varphi'(-a) = \cot\alpha$. Similarly, $\partial_{\tilde{\tau}} v(a,y) = 0$ for all $y \in (-a\cot\gamma_a, a\cot\gamma_a)$.

At the "bottom" of the boundary of Σ_a, one has $v(x, -a\cot\gamma_a) \geq 0$ for all $x \in [-a,a]$. At the "top" of the boundary of Σ_a, $v(x, a\cot\gamma_a) = u^\varepsilon(a\cot\gamma_a - \varphi(x))$ for all $x \in [-a,a]$ and

$$\forall x \in [-a,a], \quad |\varphi(x)| \leq a\cot\alpha - \frac{3}{8}\beta\cot\alpha \leq a\cot\alpha.$$

Since $(\cot\gamma_a - \cot\alpha)a \to +\infty$ as $a \to +\infty$ and since $u^\varepsilon(+\infty) = 1+\varepsilon$, it then follows that there exists a real $a_2(\varepsilon) > \beta$ such that if $a \geq a_2(\varepsilon)$ then $v(x, a\cot\gamma_a) > 1$ for all $x \in [-a,a]$.

Let us now choose $a \geq a_2(\varepsilon)$. The function v is a supersolution of problem (2.3). With the same arguments as in Lemma 2.2, we finally conclude that $v > u_c$ in $[-a,a] \times (-a\cot\gamma_a, a\cot\gamma_a)$. In particular, $u_c < v$ in $\{y = -|x|\cot\alpha,\ |x| \leq a\}$ since $0 < \gamma_a < \alpha$. As a consequence,

$$\max_{\substack{y=-\cot\alpha\ |x| \\ |x|\leq a}} u_c < \max_{\substack{y=-\cot\alpha\ |x| \\ |x|\leq a}} v \ = \ \max_{|x|\leq a} u^\varepsilon(-\cot\alpha\ |x| - \varphi(x)) = u_\varepsilon(0) = \theta. \qquad \Box$$

We complete this section with the following proposition.

Proposition 2.7. *If ε and $\eta > 0$ are small enough, then there is a real $a_0(\eta,\varepsilon) \geq A_0$ such that, for any $a \geq a_0(\eta,\varepsilon)$, problem* (2.1)–(2.2) *has a unique solution (c_a, u_a). Furthermore, one has*

$$c_0^\eta/\sin\alpha \leq c_a \leq \tilde{c}_0^\varepsilon/\sin^2\alpha.$$

Proof. Proposition 2.7 is an immediate consequence of Lemmas 2.4, 2.5, and 2.6. Indeed, let us choose $\varepsilon > 0$ and $\eta > 0$ small enough and take $a_0(\eta,\varepsilon) = \max\,(a_1(\eta), a_2(\varepsilon))$: for $a \geq a_0(\eta,\varepsilon)$, if $c < c_0^\eta/\sin\alpha$, then $\max_{\substack{y=-\cot\alpha\ |x| \\ |x|\leq a}} u_c > \theta$ from Lemma 2.5 and if $c > \tilde{c}_0^\varepsilon/\sin^2\alpha$, then $\max_{\substack{y=-\cot\alpha\ |x| \\ |x|\leq a}} u_c < \theta$ from Lemma 2.6. From Lemma 2.4, the functions u_c are continuously increasing with respect to c. Hence, problem (2.1)–(2.2) has a unique solution (c_a, u_a) and $c_0^\eta/\sin\alpha \leq c_a \leq \tilde{c}_0^\varepsilon/\sin^2\alpha$. $\Box$

2.2. Monotonicity properties of the solutions u_a. From Proposition 2.7, we assume from now on that a is large enough ($a \geq a(\eta_0, \varepsilon_0)$, where $\eta_0 > 0$, $\varepsilon_0 > 0$ are small enough) such that (2.1)–(2.2) has a unique solution (c_a, u_a). When there is no ambiguity, we call this solution (c,u). Set $\Sigma_a^- = (-a,0) \times (-a\cot\gamma_a, a\cot\gamma_a)$ and $\Sigma_a^+ = (0,a) \times (-a\cot\gamma_a, a\cot\gamma_a)$. Remember that C_i $(i = 1,\ldots,4)$ are the four corners of the rectangle Σ_a.

Proposition 2.8. *For a large enough, the unique solution (c_a, u_a) of* (2.1)–(2.2) *is such that*

(i) *for any $\rho = (\cos\beta, \sin\beta)$ with $\pi/2 - \alpha \leq \beta \leq \pi$, one has $\partial_\rho u \geq 0$ in $\overline{\Sigma_a^-} \setminus \{C_1, C_3\}$;*

(ii) *for any $\rho = (\cos\beta, \sin\beta)$ with $0 \leq \beta \leq \pi/2 + \alpha$, one has $\partial_\rho u \geq 0$ in $\overline{\Sigma_a^+} \setminus \{C_2, C_4\}$.*

From this proposition we immediately get the following corollary.

Corollary 2.9. (i) *The function u is nonincreasing with respect to x in $\overline{\Sigma_a^-}$ and nondecreasing with respect to x in $\overline{\Sigma_a^+}$.*

(ii) *For any nonzero vector $\rho \in \overline{\mathcal{C}(\vec{e}_2, \alpha)}$, one has*

$$\partial_\rho u \geq 0 \quad \text{in} \quad \tilde{\Sigma}_a = \overline{\Sigma_a} \backslash \{C_1, C_2, C_3, C_4\}.$$

Proof of Proposition 2.8. By symmetry with respect to x and by continuity, it is sufficient to prove that $\partial_\rho u \geq 0$ in Σ_a^- for any vector $\rho = (\cos\beta, \sin\beta)$ such that $\pi/2 - \alpha < \beta < \pi$. Let ρ be such a vector.

Let us temporarily consider the case where the function f is of class C^1 in $[0,1]$. Let $z = (x,y)$ be the generic notation for the points of $\overline{\Sigma_a}$. For $\varepsilon > 0$ small enough, we are going to compare the functions $u(z)$ and $u(z+\varepsilon\rho)$ in the rectangular domain $R_\varepsilon = \overline{\Sigma_a^- \cap (\Sigma_a^- - \varepsilon\rho)}$ (see Figure 3).

Let us first show that

$$u(z) < u(z+\varepsilon\rho) \quad \text{on } \partial R_\varepsilon \tag{2.9}$$

for ε small enough. Indeed, consider first the "top" and "bottom" boundaries of R_ε. Set $\vec{e}_1 = (1,0)$. If $\rho \cdot \vec{e}_1 > 0$ (as drawn in Figure 3), then those parts of ∂R_ε are $[-a, -\varepsilon\rho\cdot\vec{e}_1] \times \{-a\cot\gamma\}$ and $[-a, -\varepsilon\rho\cdot\vec{e}_1] \times \{a\cot\gamma - \varepsilon\rho\cdot\vec{e}_2\}$. Since $\rho\cdot\vec{e}_2 > 0$, inequality (2.9) is satisfied there because $u = 0$ (resp., $u = 1$) on $[-a,a] \times \{-a\cot\gamma\}$ (resp., $[-a,a] \times \{a\cot\gamma\}$) and because $0 < u < 1$ in $[-a,a] \times (-a\cot\gamma, a\cot\gamma)$. The other case $\rho\cdot\vec{e}_1 \geq 0$ can be treated similarly.

On the other hand, on $\{0\} \times [-a\cot\gamma, a\cot\gamma]$, we have $\partial_y u > 0$ from Lemma 2.2 (remember that f is assumed here to be of class C^1) and $\partial_x u = 0$ since u is symmetric with respect to x (from Corollary 2.3). Hence, $\partial_\rho u > 0$ on the compact set $\{0\} \times [-a\cot\gamma, a\cot\gamma]$. Since the function $\partial_\rho u$ is uniformly continuous in a neighborhood of $\{0\} \times [-a\cot\gamma, a\cot\gamma]$, it follows from the finite increment theorem that there exists

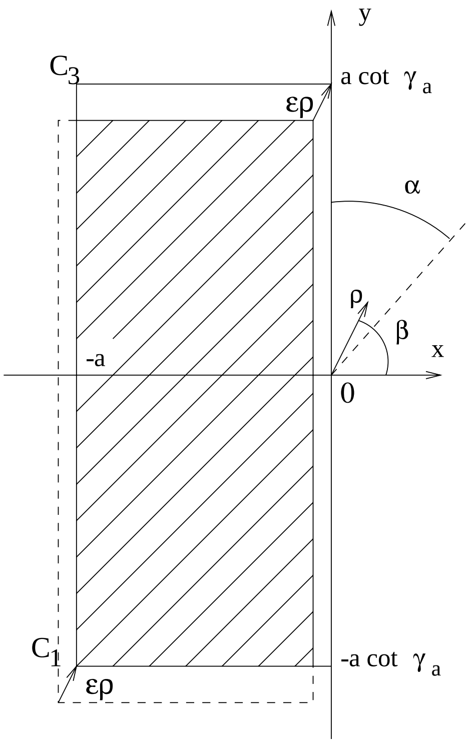


FIG. 3. *The rectangle R_ε.*

a real $\tilde{\varepsilon} > 0$ such that, if $0 < \varepsilon < \tilde{\varepsilon}$, then (2.9) is true on the right-hand side boundary of R_ε, namely, $\{-\varepsilon\rho \cdot \vec{e}_1\} \times [-a\cot\gamma, a\cot\gamma - \varepsilon\rho\cdot\vec{e}_2]$ if $\rho\cdot\vec{e}_1 \geq 0$ (as in Figure 3) or $\{0\} \times [-a\cot\gamma, a\cot\gamma - \varepsilon\rho\cdot\vec{e}_2]$ if $\rho\cdot\vec{e}_1 \leq 0$.

We now have to deal with the behavior of the function u on the left-hand boundary of R_ε and especially around the corners C_1 and C_3. We shall use the following lemma (notice that in this lemma the function f does not need to be of class C^1 in $[0,1]$).

LEMMA 2.10. *For each $i = 1$ or 3, there exist a neighborhood V_i of C_i and a real $\varepsilon_i > 0$ such that*

$$\left(0 < \varepsilon < \varepsilon_i \text{ and } z,\ z + \varepsilon\rho\ \in V_i \cap \overline{\Sigma_a}\right) \implies (u(z) < u(z + \varepsilon\rho)).$$

This technical lemma is proved in section 5.

End of the proof of Proposition 2.8. For any point $z = (-a, y_0)$ on the left-hand boundary $\{-a\} \times (-a\cot\gamma, a\cot\gamma)$ of Σ_a, we have $\partial_\tau u = 0$ and $\partial_y u > 0$ from Lemma 2.2. Since $\tau = (-\sin\alpha, -\cos\alpha)$ and $\rho = (\cos\beta, \sin\beta)$ with $\pi/2 - \alpha < \beta < \pi$, it follows that $\partial_\rho u > 0$. Since u is of class C^1 near the point z, there exists a neighborhood V_z of z such that $\partial_\rho u(x,y) > 0$ for any $(x,y) \in V_z \cap \overline{\Sigma_a}$. Hence, from the finite increment theorem, there exists a real $\varepsilon_z > 0$ such that if $0 < \varepsilon < \varepsilon_z$ and if the point $z + \varepsilon\rho$ is in $\overline{V_z} \cap \overline{\Sigma_a}$, then

$$u(z) < u(z + \varepsilon\rho).$$

Without any restriction, the neighborhoods V_1 and V_3 of C_1 and C_3, which are given in Lemma 2.10, can be replaced with two open balls $B(C_i, \delta_i)$ centered on the points C_i and with radii δ_i ($i = 1$ or 3). Since $\{-a\} \times [-a\cot\gamma + \delta_1, a\cot\gamma - \delta_3]$ is a compact set, there exists a real $\overline{\varepsilon} > 0$ such that, if $0 < \varepsilon < \overline{\varepsilon}$, if $z = (x,y)$ where $y \in [-a\cot\gamma + \delta_1, a\cot\gamma - \delta_3]$, and $x = -a$ in the case $\rho\cdot\vec{e}_1 \geq 0$ (resp., $x = -a - \varepsilon\rho\cdot\vec{e}_1$ in the case $\rho\cdot\vec{e}_1 < 0$), then $z,\ z + \varepsilon\rho \in R_\varepsilon$ and

$$u(z) < u(z + \varepsilon\rho).$$

From Lemma 2.10, we conclude that, if $0 < \varepsilon < \min(\varepsilon_1, \varepsilon_3, \overline{\varepsilon})$, then (2.9) is true on the left-hand boundary of R_ε, namely, on $\{-a - \varepsilon\rho\cdot\vec{e}_1\} \times [-a\cot\gamma, a\cot\gamma - \varepsilon\rho\cdot\vec{e}_2]$ or $\{-a\} \times [-a\cot\gamma, a\cot\gamma - \varepsilon\rho\cdot\vec{e}_2]$ according to the sign of $\rho\cdot\vec{e}_1$.

Finally, we set $\varepsilon_0 = \min(\tilde{\varepsilon}, \varepsilon_1, \varepsilon_3, \overline{\varepsilon})$ (remember that $\tilde{\varepsilon}$ has been defined just before Lemma 2.10). For any $\varepsilon \in (0, \varepsilon_0)$ and for any $z \in \partial R_\varepsilon$, the points z and $z + \varepsilon\rho$ are in $\overline{\Sigma_a}$ and we have $u(z) < u(z + \varepsilon\rho)$. Next, as in the proof of Lemma 2.2, that is to say by using a sliding method along the direction $\vec{e}_2$ and the fact that u is increasing with respect to y, we find that

$$u(z) < u(z + \varepsilon\rho) \text{ in } R_\varepsilon.$$

This completes the proof of Proposition 2.8 in the case where the function f is of class C^1 in $[0,1]$.

If f is not of class C^1 in $[0,1]$, we can however approximate it by a sequence of functions f_n of class C^1 which are such that $\|f'_n\|_{L^\infty([0,1])} \leq C$, $\|f - f_n\|_{L^\infty([0,1])} \to 0$ as $n \to +\infty$ and which satisfy (1.4) with ignition temperature $\theta_n \to \theta$ as $n \to +\infty$. Under the notation of Lemmas 2.5 and 2.6, there exist two positive reals ε_1 and η_1 such that, for n large enough, we have $f_{\eta_1} \leq f_n \leq f^{\varepsilon_1}$, whence $f_{\eta_1} \leq f \leq f^{\varepsilon_1}$ by taking the limit $n \to +\infty$. Thus, as in the proof of Proposition 2.7, for n large enough and for $a \geq \max(a_1(\eta_1), a_2(\varepsilon_1))$, we get that there exists a unique solution (c_n, u_n) of

(2.1)–(2.2) with the source term f_n as well as a unique solution (c_a, u_a) of (2.1)–(2.2) with the source term f. Furthermore, one has $c_0^{\eta_1}/\sin\alpha \le c_n \le \tilde{c}_0^{\varepsilon_1}/\sin^2\alpha$.

Choose any $a \ge \max(a_1(\eta_1), a_2(\varepsilon_1))$. First of all, up to extraction of some subsequence, we can assume that $c_n \to \tilde{c} \in \mathbb{R}$. From the standard elliptic estimates up to the boundary, we can extract a subsequence $u_{n'}$ which approaches a solution u of (2.4) with the speed $\tilde{c}$ in the spaces $W^{2,p}_{loc}(\tilde{\Sigma}_a) \cap C_{loc}(\tilde{\Sigma}_a)$. Furthermore, for each $i \in \{1,\ldots,4\}$, there exists a function $\overline{v}_i$ defined in a neighborhood V_i of the corner C_i such that $\overline{v}_i(C_i) = 0$ and, for all n' large enough,

$$\begin{array}{lrl} \text{if } i = 1 \text{ or } 2, & u_{n'}(x,y) & \le \overline{v}_i(x,y) \\ \text{if } i = 3 \text{ or } 4, & 1 - u_{n'}(x,y) & \le \overline{v}_i(x,y) \end{array} \quad \text{in } V_i \cap \overline{\Sigma_a} \tag{2.10}$$

(see Remark 5.2). As a consequence, the function $\tilde{u}$ can be extended by continuity at the four corners C_i. Hence, $\tilde{u}$ is the unique solution of (2.3) with the speed $\tilde{c}$. On the other hand, by passage to the limit $n' \to \infty$, the statements of Proposition 2.8 hold good for the function $\tilde{u}$. In particular, it follows that $\tilde{u}$ fulfills (2.2). Finally, from Lemma 2.4, we conclude that $(\tilde{c}, \tilde{u}) = (c_a, u_a)$. This completes the proof of Proposition 2.8. □

3. Passage to the limit in the whole plane. In the previous section, we proved the existence and the uniqueness of a solution (c_a, u_a) to problem (2.1)–(2.2) for a large enough. Moreover, we found several a priori bounds for the speeds c_a as well as a priori monotonicity properties for the functions u_a. We are now going to pass to the limit $a \to \infty$.

PROPOSITION 3.1. *There exists a sequence $a_n \to \infty$, a real c, and a function u such that $c_{a_n} \to c$ in $\mathbb{R}$ and $u_{a_n} \to u$ in $W^{2,p}_{loc}(\mathbb{R}^2)$ for all $p > 1$. Furthermore, the real c is such that*

$$\frac{c_0}{\sin\alpha} \le c \le \frac{c_0}{\sin^2\alpha}$$

and the function u satisfies

$$\Delta u - c\partial_y u + f(u) = 0 \textit{ in } \mathbb{R}^2, \tag{3.1}$$

$$0 < u < 1 \textit{ in } \mathbb{R}^2,$$

$$\forall (x,y) \in \mathbb{R}^2, \quad u(x,y) = u(-x,y),$$

$$\max_{\substack{y \le -\cot\alpha\ |x| \\ x \in \mathbb{R}}} u = u(0,0) = \theta, \tag{3.2}$$

$$\begin{cases} \forall \rho = (\cos\beta, \sin\beta) \textit{ such that } \pi/2 - \alpha \le \beta \le \pi, & \partial_\rho u(x,y) \ge 0 \textit{ if } x \le 0, \\ \forall \rho = (\cos\beta, \sin\beta) \textit{ such that } 0 \le \beta \le \pi/2 + \alpha, & \partial_\rho u(x,y) \ge 0 \textit{ if } x \ge 0. \end{cases} \tag{3.3}$$

COROLLARY 3.2. *For all $\rho = (\cos\beta, \sin\beta)$ with $\pi/2 - \alpha \le \beta \le \pi/2 + \alpha$, one has*

$$\partial_\rho u \ge 0 \quad \textit{in } \mathbb{R}^2.$$

Proof of Proposition 3.1. Under the notation of Proposition 2.7, choose $\varepsilon = \eta = 1/n$ where the integer n is large enough and set $a_n = a_0(1/n, 1/n)$. For n large enough, problem (2.1)–(2.2) has a unique solution (c_n, u_n) in Σ_{a_n} and one has $c_0^{1/n}/\sin\alpha \le c_n \le \tilde{c}_0^{1/n}/\sin^2\alpha$.

From the results of [9], we have $c_0^{1/n}$ and $\tilde{c}_0^{1/n} \to c_0$ as $n \to \infty$. Hence there exists a subsequence, that is still called (c_n), such that $c_n \to c \in [c_0/\sin\alpha, c_0/\sin^2\alpha]$. For any compact set K of $\mathbb{R}^2$, from the standard elliptic estimates, the sequence (u_{a_n}) is bounded in $W^{2,p}(K)$ (for a_n large enough such that $\overline{\Sigma_{a_n}} \subset \overset{o}{K}$). Hence, from the diagonal extraction process, there exists a subsequence that is still called (u_{a_n}) and a function u such that $u_{a_n} \to u$ in $W^{2,p}_{loc}(\mathbb{R}^2)$ for all $p > 1$. The function u satisfies (3.1). From the Sobolev injections and since f is Lipschitz continuous, the function u is in $C^{2,\mu}_{loc}(\mathbb{R}^2)$ for all $0 \le \mu < 1$.

Since $u(0,0) = \lim\ u_n(0,0) = \theta$ and since $0 \le u \le 1$, the strong maximum principle implies that $0 < u < 1$ in $\mathbb{R}^2$. The symmetry of u with respect to x derives from the symmetry of u_n. The assertions (3.3) come from Proposition 2.8. Together with (2.2), they yield the normalization condition (3.2). □

3.1. Exponential decay properties. For any $z = (x,y) \in \mathbb{R}^2$, let us define

$$T_z = (-|x|,|x|) \times (-\infty, y) \ \cup\ \mathcal{C}((x,y), -\vec{e}_2, \alpha) \ \cup\ \mathcal{C}((-x,y), -\vec{e}_2, \alpha).$$

Proposition 3.3. *Let x_0 be in $\mathbb{R}$.*

(i) *There exists a real $y_0 \in [-|x_0| \cot\alpha, 0]$ such that $u(x_0,y_0) = \theta$.*

(ii) *Set $z_0 = (x_0,y_0)$. The following exponential decay holds in $\overline{T_{z_0}}$:*

$$\forall z = (x,y) \in \overline{T_{z_0}}, \quad u(z) \ \ \le 2\theta e^{-c\sin\alpha\cos\alpha\ |x_0|}\cosh(c\sin\alpha\cos\alpha\ x)e^{c\sin^2\alpha\ (y-y_0)} + \theta e^{c(y-y_0)}. \tag{3.4}$$

(iii) *A similar estimate is true in $\overline{\mathcal{C}(z_0, -\vec{e}_2, \alpha)}$. Namely, for all $\pi/2 - \alpha \le \varphi \le \pi/2 + \alpha$ and $\rho = (\cos\varphi, -\sin\varphi)$, we have*

$$\forall \lambda \ge 0, \quad u(z_0 + \lambda\rho) \le 2\theta\cosh(c\lambda\sin\alpha\cos\alpha\cos\varphi)\ e^{-c\lambda\sin^2\alpha\sin\varphi}. \tag{3.5}$$

Remark 3.4. *By taking $z_0 = (0,0)$ and $\vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha)$ in* (3.5)*, it follows that the function u fulfills* (1.2) *and* (1.8).

Corollary 3.5. *The function u is increasing in y.*

Proof. From Corollary 3.2, we know that $u(x,y)$ is nondecreasing in y. Suppose that $u(x_0,y_0) = u(x_0,y_0')$ where $x_0 \in \mathbb{R}$ and $y_0 < y_0'$. It follows that u is equal to a constant u_0 in $\overline{\mathcal{C}((x_0,y_0), \vec{e}_2, \alpha)} \cap \overline{\mathcal{C}((x_0,y_0'), -\vec{e}_2, \alpha)}$. This constant u_0 is then a zero of the function f. Since $0 < u < 1$ in $\mathbb{R}^2$ and $f > 0$ on $(\theta, 1)$, we get $u_0 \in (0, \theta]$. The monotonicity properties imply that $u \le u_0$ in the cone $\mathcal{C} = \mathcal{C}((x_0,y_0'), -\vec{e}_2, \alpha)$ and that the function u satisfies

$$\Delta u - c\partial_y u = 0 \ \text{ in } \mathcal{C}.$$

In $\mathcal{C}$, the function u reaches its maximum u_0 at an interior point, for instance, $(x_0, (y_0 + y_0')/2)$. From the strong maximum principle, u is then equal to u_0 in $\overline{\mathcal{C}}$. This is impossible because $u(x_0,y) \to 0$ as $y \to -\infty$ from inequality (3.5). □

Proof of Proposition 3.3. From the symmetry of u with respect to x, we may suppose that $x_0 \ge 0$. Let now $a > x_0$. By Proposition 2.8, we have $u_a(x_0, 0) \ge \theta$ and $u_a(x_0, -x_0\cot\alpha) \le \theta$. Since u_a is continuous, there exists a real y_a in $[-x_0\cot\alpha, 0]$ such that $u_a(x_0,y_a) = \theta$. Since the y_a are bounded and since the functions u_a approach u in $C^1_{loc}(\mathbb{R}^2)$ (for a certain sequence $a \to +\infty$), then there exists a real y_0 in $[-x_0\cot\alpha, 0]$ such that $y_a \to y_0$ (for a sequence $a \to \infty$) and $u(x_0,y_0) = \theta$. This yields the assertion (i) of Proposition 3.3.

Let $z_0 = (x_0,y_0)$. Let us now consider the open trapezium D_a whose vertices are the four points $C_1 = (-a, -a\cot\gamma_a)$, $S_1 = (-x_0, y_a)$, $S_2 = (x_0, y_a)$, and $C_2 =$

$(a, -a\cot\gamma_a)$. The angles between $-\vec{e_2}$ and each side $[S_1, C_1]$ and $[S_2, C_2]$ are equal and, since $y_a \geq -x_0\cot\alpha \geq -x_0\cot\gamma_a$, they are not larger than γ_a and, a fortiori, they are less than α. Hence, from Proposition 2.8 we have

$$u_a \leq \theta \ \text{ in } \overline{D_a}$$

and

$$\Delta u_a - c_a\partial_y u_a = 0 \ \text{ in } D_a.$$

We are now going to compare u_a with the sum of three exponential functions in D_a. Choose any point $z_1 = (x_1, y_1)$ in the open set T_{z_0}. Since $y_a \to y_0$ and $\gamma_a \to \alpha$, there exists a positive real a_0 such that $z_1 \in D_a$ for all $a \geq a_0$. Let c' be a real in $(0, c\sin\alpha)$ − notice that this is possible since $\sin\alpha > 0$ and $c\sin\alpha \geq c_0 > 0$. Let us set $r_a = 1/\sqrt{(a\cot\gamma_a + y_a)^2 + (-a+x_0)^2}$ and define

$$w_a(x,y) = f_1(x,y) + f_2(x,y) + f_3(x,y),$$

where

$$\begin{cases} f_1(x,y) & = \theta e^{-c'r_a((a\cot\gamma_a+y_a)(x+x_0)+(x_0-a)(y-y_a))}, \\ f_2(x,y) & = \theta e^{-c'r_a(-(a\cot\gamma_a+y_a)(x-x_0)+(x_0-a)(y-y_a))}, \\ f_3(x,y) & = \theta e^{c'/\sin\alpha\ (y-y_a)}. \end{cases}$$

In particular, we have $w_a \geq \theta \geq u_a$ on ∂D_a. Moreover, a straightforward calculation gives

$$\Delta w_a - c_a\partial_y w_a = c'(c' - c_a r_a(a - x_0))(f_1 + f_2) + \frac{c'}{\sin^2\alpha}\ (c' - c_a\sin\alpha)f_3.$$

Since $c' > 0$ and since $c_a \to c > c'/\sin\alpha$, $\ r_a(a - x_0) \to \sin\alpha$ as $a \to \infty$, it follows that

$$\Delta w_a - c_a\partial_y w_a < 0 \ \text{ in } D_a$$

for a large enough. From the maximum principle, we deduce that $u_a < w_a$ in D_a. By passing to the limit $a \to \infty$, we obtain

$$\begin{aligned} u(x_1,y_1) \ & \leq \theta e^{-c'[\cos\alpha(x_1+x_0)-\sin\alpha(y_1-y_0)]} \\ & \quad + \theta e^{-c'[-\cos\alpha(x_1-x_0)-\sin\alpha(y_1-y_0)]} + \theta e^{c'/\sin\alpha\ (y_1-y_0)}. \end{aligned}$$

Since this is true for any $c' < c\sin\alpha$, we can pass to the limit $c' \to c\sin\alpha$ and we get

$$u(x_1,y_1) \leq 2\theta\cosh(c\sin\alpha\cos\alpha\ x_1)\ e^{c\sin^2\alpha\ (y_1-y_0) - c\sin\alpha\cos\alpha\ x_0} + \theta e^{c(y_1-y_0)}.$$

This can be extended by continuity in $\overline{T_{z_0}}$. This gives assertion (ii) of Proposition 3.3.

In the same way, we could prove that for any $x_0 \geq 0$,

$$u(x,y) \leq 2\theta\cosh(c\sin\alpha\cos\alpha\ (x - x_0))e^{c\sin^2\alpha(y-y_0)} \ \text{ in } \overline{\mathcal{C}(z_0, -\vec{e_2}, \alpha)}$$

by comparing the function u_a with the sum of two suitable exponential functions in the triangles whose vertices are $S_1 = (-a + 2x_0, -a\cot\gamma_a)$, $S_2 = (x_0, y_0)$, and $S_3 = (a, -a\cot\gamma_a)$. This corresponds to assertion (iii) of Proposition 3.3. The case $x_0 \leq 0$ can be treated by symmetry. □

3.2. Estimating the speed c: Proof of formula (1.7). Consider now a sequence $x_n \to -\infty$ and, for any x_n, let y_n be the unique real such that $u(x_n, y_n) = \theta$. One has $x_n \cot \alpha \leq y_n \leq 0$. Move the origin at the point (x_n, y_n) and consider the functions

$$v_n(x,y) = u(x + x_n, y + y_n) \text{ in } \mathbb{R}^2.$$

From the standard elliptic estimates and the Sobolev injections, the functions v_n are bounded in $W^{2,p}_{loc}(\mathbb{R}^2)$ for all $1 < p < \infty$ and approach, up to extraction of some subsequence, a function $v \in \bigcap_{p>1} W^{2,p}_{loc}(\mathbb{R}^2)$, such that

$$\begin{cases} \Delta v - c\partial_y v + f(v) = 0 & \text{in } \mathbb{R}^2, \\ v(0,0) = \theta. \end{cases} \tag{3.6}$$

The function v has the following monotonicity properties.

LEMMA 3.6. *For any $\rho = (\cos\varphi, -\sin\varphi)$ such that $0 \leq \varphi \leq \pi/2 + \alpha$, one has the following:*

(i) *the function v is nonincreasing in the direction ρ;*

(ii) *it also holds that*

$$\forall \lambda \geq 0, \quad v(\lambda\rho) \leq \theta e^{-c\lambda \sin\alpha \cos(\alpha - \varphi)} + \theta e^{-c\lambda \sin\varphi}. \tag{3.7}$$

Proof. Let ρ be as in the lemma above. Let $z = (x,y)$ be any point in $\mathbb{R}^2$ and let $\lambda > 0$. Consider both points z and $z + \lambda\rho$. Since $x_n \to -\infty$, we have $x + x_n \leq 0$ and $x + x_n + \lambda\cos\varphi \leq 0$ for n large enough. From (3.3), we have, for n large enough,

$$v_n(z) = u(x + x_n, y + y_n) \geq u(x + x_n + \lambda\cos\varphi, y + y_n - \lambda\sin\varphi) = v_n(z + \lambda\rho).$$

By taking the limit $n \to \infty$, it follows that $v(z) \geq v(z + \lambda\rho)$. This gives the assertion (i).

Consider the set

$$T_n = (-|x_n|, |x_n|) \times (-\infty, y_n) \ \cup \ \mathcal{C}((x_n, y_n), -\vec{e}_2, \alpha) \ \cup \ \mathcal{C}((-x_n, y_n), -\vec{e}_2, \alpha).$$

Under the notation of section 3.1, we have $T_n = T_{z_n = (x_n, y_n)}$. Since $x_n \to -\infty$, the points $(x_n, y_n) + \lambda\rho$ are in $\overline{T_n}$ for n large enough. Hence, inequality (3.4) implies that

$$\begin{aligned} v_n(\lambda\rho) \quad &\leq 2\theta e^{-c|x_n| \sin\alpha\cos\alpha} \cosh(c \sin\alpha \cos\alpha \ (x_n + \lambda\cos\varphi)) e^{-c\lambda \sin^2\alpha \sin\varphi} \\ &+ \theta e^{-c\lambda\sin\varphi}. \end{aligned}$$

Since $x_n \to -\infty$, we obtain at the limit $n \to \infty$

$$v(\lambda\rho) \leq \theta e^{-c\lambda \sin\alpha\cos\alpha\cos\varphi} e^{-c\lambda\sin^2\alpha\sin\varphi} + \theta e^{-c\lambda\sin\varphi}.$$

This completes the proof of Lemma 3.6. □

PROPOSITION 3.7. *The speed c is equal to $c_0/\sin\alpha$.*

Proof. From (1.7), we already know that $c_0/\sin\alpha \leq c \leq c_0/\sin^2\alpha$. Let us suppose that $c > c_0/\sin\alpha$.

First step: Construction of a supersolution. As in the proof of Lemma 2.6, we use the same functions $f^\varepsilon \geq f$ such that $f^\varepsilon \equiv 0$ on $[0, \theta - \varepsilon] \cup \{1 + \varepsilon\}$, $f^\varepsilon > 0$ on

$(\theta-\varepsilon,1+\varepsilon)$, and $f^\varepsilon\to f$ as $\varepsilon\to 0$ uniformly in $[0,1]$. For each $\varepsilon>0$, there exists a unique solution $(\overline{c}_0^\varepsilon,U^\varepsilon)$ of

$$\left\{\begin{array}{rl} (U^\varepsilon)''-\overline{c}_0^\varepsilon(U^\varepsilon)'+f^\varepsilon(U^\varepsilon) & =0 \qquad \text{in } \mathbb{R},\\ U^\varepsilon(-\infty)=\varepsilon,\ U^\varepsilon(0)=\theta,\ U^\varepsilon(+\infty) & =1+\varepsilon. \end{array}\right. \tag{3.8}$$

From the results in [9], we have $\overline{c}_0^\varepsilon\to c_0$ as $\varepsilon\to 0$. Now choose $\varepsilon>0$ such that

$$c>\overline{c}_0^\varepsilon/\sin\alpha$$

and denote by U the function U^ε.

Let us consider the new variables

$$X=y\cos\alpha+x\sin\alpha \quad\text{ and }\quad Y=y\sin\alpha-x\cos\alpha.$$

The variables (X,Y) are obtained from (x,y) by a rotation of angle $\pi/2-\alpha$ around the origin.

We are looking for a supersolution of (3.6) of the type

$$w(x,y)=U(Y-\phi(X)).$$

For such a function w, we have

$$\Delta w-c\partial_y w+f(w)=A(X)U'(Y-\phi(X))+f(U)-f_\varepsilon(U)-\phi'^2 f_\varepsilon(U), \tag{3.9}$$

where

$$A(X)=\overline{c}_0^\varepsilon(1+\phi'^2)-\phi''-c(\sin\alpha-\cos\alpha\ \phi').$$

Since $f_\varepsilon\geq f\geq 0$ and $U'>0$, in order to make the right-hand side of (3.9) nonpositive, it is sufficient to choose a function ϕ in such a way that $A(X)\leq 0$. Let ϕ be defined by

$$\phi(X)=-\frac{1}{c\sin\alpha}\ln(e^{-c\sin\alpha\tan\beta\ X}+e^{c\sin\alpha\cot(\alpha-\beta)X}),$$

where $\beta>0$ shall be chosen later. Set $\delta=\cot(\alpha-\beta)+\tan\beta$. It is easy to check that

$$A(X)=\frac{1}{(1+e^{c\sin\alpha\delta X})^2}\left[B(\beta)e^{2c\sin\alpha\delta X}+C(\beta)e^{c\sin\alpha\delta X}+D(\beta)\right],$$

where

$$\left\{\begin{array}{rcl} B(\beta) & = & \overline{c}_0^\varepsilon-c\sin\alpha-c\cos\alpha\cot(\alpha-\beta)+\overline{c}_0^\varepsilon\cot^2(\alpha-\beta),\\ C(\beta) & = & 2(\overline{c}_0^\varepsilon-c\sin\alpha)-c\cos\alpha\cot(\alpha-\beta)\\ & & +c\cos\alpha\tan\beta-2\overline{c}_0^\varepsilon\tan\beta\cot(\alpha-\beta)+c\sin\alpha\ \delta^2,\\ D(\beta) & = & \overline{c}_0^\varepsilon-c\sin\alpha+c\cos\alpha\tan\beta+\overline{c}_0^\varepsilon\tan^2\beta. \end{array}\right.$$

As $\beta\to 0$, we have $B(\beta)\to\overline{c}_0^\varepsilon/\sin^2\alpha-c/\sin\alpha<0$, $C(\beta)\to 2(\overline{c}_0^\varepsilon-c\sin\alpha)<0$, and $D(\beta)\to\overline{c}_0^\varepsilon-c\sin\alpha<0$. Hence, we can choose $\beta\in(0,\alpha)$ small enough such that $B(\beta)$, $C(\beta)$, $D(\beta)<0$.

Let β be chosen as above. The function $w(x,y)$ is then a supersolution of (3.6) in the sense that

$$\Delta w-c\partial_y w+f(w)<0\ \text{ in } \mathbb{R}^2. \tag{3.10}$$

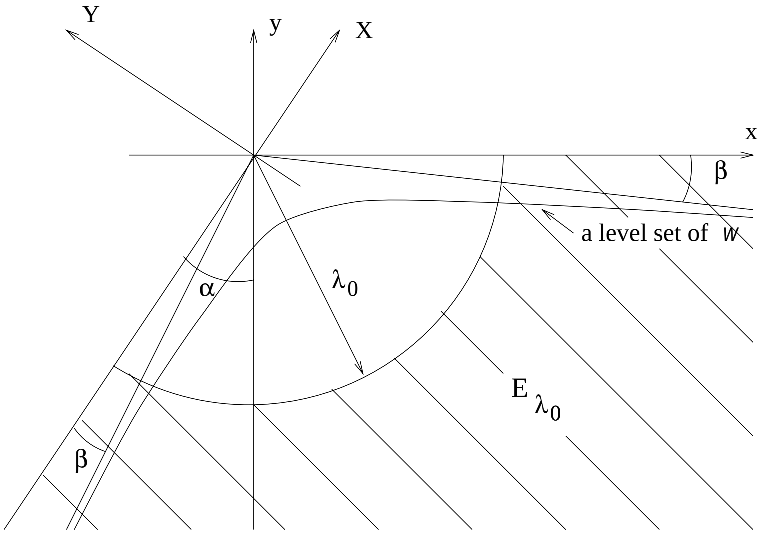


FIG. 4. *The set E_{λ_0}.*

Second step: Initialization of a sliding method. For any λ_0, we set

$$E_{\lambda_0} = \{z = (\lambda\cos\varphi, -\lambda\sin\varphi) \in \mathbb{R}^2,\ 0 \le \varphi \le \pi/2 + \alpha,\ \lambda \ge \lambda_0\} \tag{3.11}$$

(see Figure 4).

LEMMA 3.8. *There exists $\lambda_0 > 0$ such that*

$$w > v \quad in\ E_{\lambda_0}.$$

Proof. Assume that the previous conclusion is not true. There exist then two sequences $0 \le \lambda_n \to +\infty$ and $z_n = (x_n, y_n) = (\lambda_n\cos\varphi_n,\ -\lambda_n\sin\varphi_n) \in E_{\lambda_n}$ such that $w(z_n) \le v(z_n)$.

Set $X_n = y_n\cos\alpha + x_n\sin\alpha = \lambda_n\sin(\alpha - \varphi_n)$ and $Y_n = y_n\sin\alpha - x_n\cos\alpha = -\lambda_n\cos(\alpha - \varphi_n)$. From (3.6) and Lemma 3.6 (i), it follows that $v \le \theta$ in E_{λ_0} and a fortiori in E_{λ_n} for n large enough. Hence, $w(z_n) = U(Y_n - \phi(X_n)) \le \theta$. Since U is increasing and $U(0) = \theta$, we get that $Y_n - \phi(X_n) \le 0$. On the other hand, from equation (3.8) satisfied by U, we have

$$\forall \xi \le 0, \quad U(\xi) = \varepsilon + (\theta - \varepsilon)e^{\overline{c}_0^\varepsilon \xi}.$$

Hence,

$$w(z_n) = U(Y_n - \phi(X_n)) = \varepsilon + (\theta - \varepsilon)e^{\overline{c}_0^\varepsilon (Y_n - \phi(X_n))} \le v(z_n). \tag{3.12}$$

Since $\varphi_n \in [0, \pi/2 + \alpha]$, up to extraction of some subsequence, the following two cases occur.

(i) $\varphi_n \to \varphi \in]0, \pi/2 + \alpha[$. In this case, inequality (3.7) implies that $v(z_n) \to 0$ as $n \to +\infty$, whereas the left-hand side of (3.12) is greater than the positive constant ε. Case (i) is then impossible.

(ii) $\varphi_n \to 0$ or $\pi/2 + \alpha$. Since $\beta > 0$ and since each level set of the function $Y - \phi(X)$ has two asymptotes directed by the vectors $\rho_1 = (\cos\beta, -\sin\beta)$ and $\rho_2 =$

$(\cos(\pi/2+\alpha-\beta), -\sin(\pi/2+\alpha-\beta))$, the distance between the points z_n and the half-lines $\mathbb{R}_+\rho_1$, $\mathbb{R}_+\rho_2$ necessarily approaches $+\infty$. This finally yields that $Y_n - \phi(X_n) \to +\infty$, whence $w(z_n) \to 1+\varepsilon$ as $n \to \infty$. This is ruled out by the inequality $w(z_n) \le v(z_n) < 1$.

This completes the proof of Lemma 3.8. □

Third step: The sliding method. We are now going to slide w in the Y-direction and compare it with the function v. For all $\tau \in \mathbb{R}$, we set

$$w_\tau(x,y) = U(\tau + Y - \phi(X)).$$

From Lemma 3.8, there exists a real λ_0 such that $w > v$ in E_{λ_0}, whence $w_\tau > v$ in E_{λ_0} for any $\tau \ge 0$ (remember that U is increasing).

The level set $\{Y - \phi(X) = 1+\varepsilon/2\}$ of w has two asymptotes directed by the vectors $(\cos\beta, -\sin\beta)$ and $(\cos(\pi/2+\alpha-\beta), -\sin(\pi/2+\alpha-\beta))$. Owing to the definition of E_{λ_0} and since $0<\beta$, there exists a real $\tau > 0$ such that the shifted level set $\{Y+\tau-\phi(X) = 1+\varepsilon/2\}$ in the direction Y is included in E_{λ_0}.

We now claim that

$$w_\tau > v \text{ in } \mathbb{R}^2.$$

Indeed, we already know that this is true in E_{λ_0}. But in $\mathbb{R}^2\backslash E_{\lambda_0}$, we have $w_\tau(x,y) = U(\tau+Y-\phi(X)) \ge 1+\varepsilon/2$ from the definition of τ. Hence,

$$w_\tau(x,y) \ge 1+\varepsilon/2 > v(x,y) \ \text{ in } \mathbb{R}^2\backslash E_{\lambda_0}.$$

Let us now slide w in the Y-direction. In other words, let us decrease τ and call

$$\tau^* = \inf\,\{\tau \in \mathbb{R},\ w_\tau > v \text{ in } \mathbb{R}^2\}.$$

This real is finite because $w_\tau(0,0) \to U(-\infty) = \varepsilon < \theta$ as $\tau \to -\infty$ and $v(0,0) = \theta$. Since U is increasing, we have $w_\tau > v$ for all $\tau > \tau^*$. By continuity, we find that

$$w_{\tau^*} \ge v \text{ in } \mathbb{R}^2.$$

Since the function w_{τ^*} satisfies (3.10), the nonnegative function $z = w_{\tau^*} - v$ is such that

$$\Delta z - c\partial_y z + c(x,y)z \le 0 \ \text{ in } \mathbb{R}^2$$

for some bounded function $c(x,y)$. From the strong maximum principle, one of the following two situations occurs:

(i) $w_{\tau^*} \equiv v$ in $\mathbb{R}^2$,

(ii) $w_{\tau^*} > v$ in $\mathbb{R}^2$.

Case (i) cannot occur since $w_{\tau^*} \to 1+\varepsilon$ as $Y \to +\infty$, whereas $v < 1$ in $\mathbb{R}^2$. If case (ii) occurs, let us consider an increasing sequence $\tau_n \to \tau^*$. For each n, owing to the definition of τ^*, there exists a point $(x_n,y_n) \in \mathbb{R}^2$ such that $w_{\tau_n}(x_n,y_n) \le v(x_n,y_n)$. The points (x_n,y_n) cannot be bounded; otherwise there would exist a point $(\overline{x},\overline{y}) \in \mathbb{R}^2$ such that $w_{\tau^*}(\overline{x},\overline{y}) \le v(\overline{x},\overline{y})$. The latter is impossible because of assumption (ii). Now, as in Lemma 3.8, there exists a real $\tilde{\lambda}_0$ such that $w_{\tau_0} > v$ in $E_{\tilde{\lambda}_0}$. Since the sequence (τ_n) is increasing, we have $w_{\tau_n} > v$ in $E_{\tilde{\lambda}_0}$. This implies that $(x_n,y_n) \notin E_{\tilde{\lambda}_0}$. On the other hand, since $0<\beta$ and since any level set of the function $Y-\phi(X)$ has two asymptotes directed by the vectors $\rho_1 = (\cos\beta, -\sin\beta)$ and $\rho_2 = (\cos(\pi/2+\alpha-\beta), -\sin(\pi/2+\alpha-\beta))$, it follows that $w_{\tau_n}(x_n,y_n) \to 1+\varepsilon$ as $n \to \infty$. This is impossible since $w_{\tau_n}(x_n,y_n) \le v(x_n,y_n) < 1$.

Finally, the assertion $c > c_0/\sin\alpha$ was impossible. Hence, $c = c_0/\sin\alpha$. This completes the proof of Proposition 3.7. □

3.3. Convergence of the function u to a planar wave far away from the axis of symmetry. The case $\alpha = \pi/2$ is treated separately. Indeed, in this case, from the uniqueness result in Lemma 2.2, the functions u_a only depend on y and they solve $u_a'' - c_a u_a' + f(u_a) = 0$, $u_a(-a\cot\gamma_a) = 0$, $u_a(0) = \theta$, and $u_a(a\cot\gamma_a) = 1$. From the construction given in [9], those functions u_a approach the solution $U(y)$ of (1.5) as $a \to +\infty$. This immediately yields the asymptotic limit (1.3) as well as the last assertion of Theorem 1.1.

In the case where $\alpha < \pi/2$, as in section 3.2, we again consider the function v, obtained as the limit of the functions $v_n(x,y) = u(x + x_n, y + y_n)$, where $x_n \to -\infty$ and $u(x_n, y_n) = \theta$. We know that the function v is nonincreasing in each direction $\rho = (\cos\varphi, -\sin\varphi)$ such that $0 \le \varphi \le \pi/2 + \alpha$. Furthermore, v has an exponential decay in the set $\{\lambda(\cos\varphi, -\sin\varphi),\ \lambda \ge 0, 0 \le \varphi \le \pi/2 + \alpha\}$ of the type (3.7).

Our goal is to prove that v is actually equal to the planar wave $U(Y) = U(y\sin\alpha - x\cos\alpha)$. We divide the proof into four main steps.

First step: Construction of a supersolution. We still use the variables $X = y\cos\alpha + x\sin\alpha$ and $Y = y\sin\alpha - x\cos\alpha$. In the previous section, we considered a supersolution of (3.6) of the type $w(x,y) = U^\varepsilon(Y - \phi(X))$, which had two asymptotes directed by the two vectors $\rho_1 = (\cos\beta, -\sin\beta)$ and $\rho_2 = (\cos(\pi/2+\alpha-\beta), -\sin(\pi/2+\alpha-\beta))$ ($\beta > 0$ was a small angle).

Now, consider the function w defined by

$$w(x,y) = U(Y - \phi(X)),$$

where U is the unique solution of (1.5) such that $U(0) = \theta$ and where

$$\phi(X) = -\frac{1}{c_0}\ln(1 + e^{c_0\cot\alpha\ X}).$$

Since $c = c_0/\sin\alpha$, we have

$$\Delta w - c\partial_y w + f(w) = -\phi'(X)^2 f(U(Y - \phi(X))) \le\not\equiv 0 \ \text{ in } \mathbb{R}^2. \tag{3.13}$$

Second step: Initialization of a sliding method. Let $h(X)$ be the function defined as follows:

$$h(X) = \begin{cases} 0 & \text{if } X \le 0, \\ -X\cot\alpha & \text{if } X \ge 0. \end{cases}$$

Set $E_0 = \{\lambda(\cos\varphi, -\sin\varphi),\ \lambda \ge 0, 0 \le \varphi \le \pi/2 + \alpha\} = \{Y \le h(X)\}$ (this definition is the same as (3.11)). We claim that

$$w \ge v \ \text{ in } E_0. \tag{3.14}$$

Indeed, let $(x,y) = (\lambda\cos\varphi, -\lambda\sin\varphi) \in E_0$ with $\lambda \ge 0$ and $0 \le \varphi \le \pi/2 + \alpha$. We have $X = \lambda\sin(\alpha - \varphi)$, $Y = -\lambda\cos(\alpha - \varphi)$, and

$$w(x,y) = U(-\lambda\cos(\alpha - \varphi) - \phi(\lambda\sin(\alpha - \varphi))).$$

From Lemma 3.6 (i) and since $v(0,0) = \theta$, one has $v \le \theta$ in E_0. Hence, inequality (3.14) is immediately satisfied if $w \ge \theta$. Consider now the case where $w(x,y) \le \theta$. Since $U(\xi) = \theta e^{c_0\xi}$ for $\xi \le 0$, it follows that

$$\begin{aligned} w(x,y) &= U(-\lambda\cos(\alpha - \varphi) - \phi(\lambda\sin(\alpha - \varphi))) \\ &= \theta e^{c_0(-\lambda\cos(\alpha-\varphi) + \frac{1}{c_0}\ln(1+e^{c_0\lambda\cot\alpha\sin(\alpha-\varphi)}))} \\ &= \theta(e^{-c\lambda\sin\alpha\cos(\alpha-\varphi)} + e^{-c\lambda\sin\varphi}) \\ &\ge v(x,y) \quad \text{by (3.7).} \end{aligned}$$

For any $\tau \in \mathbb{R}$, we set $w_\tau(x,y) = U(\tau + Y - \phi(X))$. Since U is increasing, we have

$$\forall \tau \geq 0, \quad w_\tau \geq v \ \text{ in } E_0. \tag{3.15}$$

On the half-line $\{Y = 0, X \leq 0\}$ of ∂E_0, we have $Y - \phi(X) = -\phi(X) \geq 0$. On the other half-line $\{Y = -\cot\alpha\ X, X \geq 0\}$ of ∂E_0, we have $Y - \phi(X) = -\cot\alpha\ X + 1/c_0 \ln(1 + e^{c_0 \cot\alpha\ X}) \geq 0$. Thus $w_\tau \geq U(\tau)$ on ∂E_0.

Since $f'_-(1) = \lim_{t\to 1,\ t<1} \frac{f(t)-f(1)}{t-1} < 0$ and $f \equiv 0$ on $[1, \infty[$, there exists a real $\varepsilon \in (0, 1-\theta)$ such that

$$(\ t \leq s \in [1-\varepsilon, 1]\ ) \implies \left(\ f(s) - f(t) \leq \frac{f'_-(1)}{2}\ (s-t) \leq 0\ \right). \tag{3.16}$$

Since U is increasing and approaches 1 at $+\infty$, there exists a real $\tau_1 \geq 0$ such that

$$\forall \tau \geq \tau_1, \quad w_\tau \geq 1 - \varepsilon \text{ on } \partial E_0. \tag{3.17}$$

Since the function w increases with respect to Y, we finally conclude from the definition of E_0 that

$$\forall \tau \geq \tau_1, \quad w_\tau \geq 1 - \varepsilon \ \text{ in } \mathbb{R}^2 \backslash E_0.$$

Lemma 3.9. *For all $\tau \geq \tau_1$, $w_\tau \geq v$ in $\mathbb{R}^2$.*

Proof. Choose any $\tau \geq \tau_1$. By (3.15) and since $\tau_1 \geq 0$, we already know that $w_\tau \geq v$ in E_0.

Let $\tilde{\Omega}_+$ be the open set $\tilde{\Omega}_+ = \mathbb{R}^2 \backslash E_0\ \cap\ \{w_\tau < v\}$. In order to prove Lemma 3.9, the only thing we still need to prove is that $\tilde{\Omega}_+$ is empty. Set $z = w_\tau - v$. From (3.6) and (3.13) we have

$$\Delta z - c\partial_y z \leq f(v) - f(w_\tau) \ \text{ in } \mathbb{R}^2.$$

In $\tilde{\Omega}_+$, the function v satisfies $1 \geq v > w_\tau \geq 1 - \varepsilon$ from (3.17). From the choice of ε (see (3.16)), we finally get

$$\Delta z - c\partial_y z + f'_-(1)/2\ z \leq 0 \ \text{ in } \tilde{\Omega}_+. \tag{3.18}$$

If $\tilde{\Omega}_+$ is not empty, define $-\delta = \inf_{\tilde{\Omega}_+} z$ (we have $-\varepsilon \leq -\delta < 0$) and consider a sequence $(x_n, y_n) \in \tilde{\Omega}_+$ such that $z(x_n, y_n) \to -\delta$ as $n \to \infty$. From the standard elliptic estimates, ∇z is bounded in $\mathbb{R}^2$. There exists then a real $r > 0$ such that the open ball $B((x_n, y_n), r)$ lies in $\tilde{\Omega}_+$ for n large enough. The functions $z_n(x,y) = z(x + x_n, y + y_n)$ approach, up to extraction of some subsequence, a function $\tilde{z}$ defined at least in $B((0,0), r)$. This function $\tilde{z}$ reaches its minimum $-\delta < 0$ at the point $(0,0)$ and it satisfies (3.18) in $B((0,0), r)$. This is clearly impossible since $f'_-(1) < 0$. Hence, $\tilde{\Omega}_+ = \emptyset$ and $w_\tau \geq v$ in $\mathbb{R}^2$ for all $\tau \geq \tau_1$. □

Third step: Sliding method. We now decrease τ and we are going to prove the following lemma.

Lemma 3.10. *There exist two reals τ^*, $\overline{Y}$ and a sequence of points (x_n, y_n) such that the coordinates (X_n, Y_n) satisfy $X_n \to -\infty$, $Y_n \to \overline{Y}$, and*

$$v_n(x,y) = v(x + x_n, y + y_n) \ \to \ U(\tau^* + \overline{Y} + Y) \ \textit{as } n \to \infty$$

in the spaces $W^{2,p}_{loc}(\mathbb{R}^2)$ *for all* $p > 1$.

Proof. Call

$$\mathcal{E} = \{\tau,\ w_\tau \geq v \text{ in } \mathbb{R}^2\}.$$

The set $\mathcal{E}$ is not empty from Lemma 3.9. Let us define

$$\tau^* = \inf\ \mathcal{E}.$$

The real τ^* is finite since $w_\tau(x, y) \to 0$ as $\tau \to -\infty$ for any $(x, y) \in \mathbb{R}^2$. By continuity with respect to τ, we have

$$w_{\tau^*} \geq v.$$

Since the function w_{τ^*} is a strict supersolution of (3.1) in the sense that it satisfies (3.13), the strong maximum principle yields that $w_{\tau^*} > v$ in $\mathbb{R}^2$.

Remember that ε satisfies (3.16). Owing to the definition of w, there exists a real $A \geq 0$ such that

$$w_{\tau^*} \geq 1 - \varepsilon/2 \ \text{ on } \{Y = h(X) + A\}. \tag{3.19}$$

Let us set $\Omega_+ = \{Y \geq h(X) + A\}$ and $\Omega_- = E_0 = \{Y \leq h(X)\}$. By (3.6) and Lemma 3.6, we have already seen that $v \leq \theta$ in Ω_-. Last, let $\mathcal{B} = \{h(X) < Y < h(X) + A\} = \mathbb{R}^2 \backslash (\Omega_+ \cup \Omega_-)$ (see Figure 5).

Comparison of $w_{\tau^*-\delta}$ *and* v *on* $\partial\Omega_+$. Since the function w is Lipschitz continuous and fulfills (3.19), we have $w_{\tau^*-\delta} \geq 1 - \varepsilon$ on $\partial\Omega_+ = \{Y = h(X) + A\}$ if $\delta \in (0, \delta_0)$ for δ_0 small enough. Two cases may occur:

(i) There exists $\delta_1 \in (0, \delta_0)$ such that $w_{\tau^*-\delta_1} > v$ on $\partial\Omega_+$.

(ii) For n large enough, there exists a point $(x_n, y_n) \in \partial\Omega_+$ such that

$$w_{\tau^*-1/n}(x_n, y_n) \leq v(x_n, y_n). \tag{3.20}$$

Study of case (i). In this case, we argue as in the proof of Lemma 3.9 and conclude that $w_{\tau^*-\delta_1} \geq v$ in Ω_+. As a consequence, for all $\delta \in [0, \delta_1]$, one has $w_{\tau^*-\delta} \geq v$ in Ω_+.

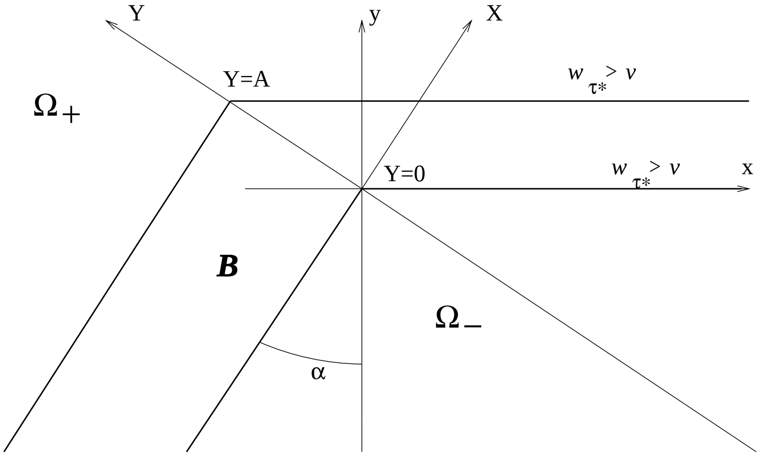


FIG. 5. *The sets* Ω_+, Ω_-, *and* $\mathcal{B}$.

Study of case (ii). In this case, the points (x_n, y_n) cannot be bounded; otherwise there exists a point $(\overline{x}, \overline{y}) \in \partial\Omega_+$ such that $w_{\tau^*}(\overline{x}, \overline{y}) = v(\overline{x}, \overline{y})$. But we have already seen that $w_{\tau^*} > v$ in $\mathbb{R}^2$. Hence one of the following situations occurs:

(ii)(a) There exists a subsequence of (x_n, y_n) such that $X_n \to -\infty$, and $Y_n = A$. We set

$$\begin{cases} w_n(x,y) &= w_{\tau^*}(x + x_n, y + y_n) \quad \text{in } \mathbb{R}^2, \\ v_n(x,y) &= v(x + x_n, y + y_n) \quad \text{in } \mathbb{R}^2. \end{cases}$$

Up to extraction of some subsequence, the functions v_n approach a solution v_∞ of (1.1) and the functions w_n approach the function $w_\infty = U(\tau^* + A + Y)$ in the spaces $W^{2,p}_{loc}(\mathbb{R}^2)$. At the limit $n \to +\infty$, we get

$$w_\infty \geq v_\infty \text{ in } \mathbb{R}^2. \tag{3.21}$$

Since the function w_τ has bounded derivatives, we conclude from (3.20) and (3.21) that $w_\infty(0,0) = v_\infty(0,0)$. Now, both functions v_∞ and w_∞ solve (1.1). From the strong maximum principle, we conclude that

$$v_\infty \equiv w_\infty = U(\tau^* + A + Y).$$

That gives the conclusion of Lemma 3.10.

(ii)(b) There exists a subsequence of (x_n, y_n) such that $x_n \to +\infty$, $y_n = A \sin\alpha$. We again normalize the functions w_{τ^*} and v as in case (ii)(a). Under the same notation as in case (ii)(a), we have $w_\infty = U((1/\sin\alpha)\ (y + A\sin\alpha) + \tau^*) \geq v_\infty$ and $w_\infty(0,0) = v_\infty(0,0)$. On the other hand, the function w_∞ is a solution of

$$\Delta w_\infty - c\partial_y w_\infty + f(w_\infty) = (1 - 1/\sin^2\alpha)\ f(U((1/\sin\alpha)\ (y + A\sin\alpha) + \tau^*)).$$

Since $\alpha < \pi/2$, the function w_∞ is then a strict supersolution of (1.1), whereas v_∞ is a solution. This is ruled out by the strong maximum principle.

As a conclusion of this part, only the cases (i) or (ii)(a) may occur and case (ii)(a) leads to the conclusion of Lemma 3.10.

Comparison of $w_{\tau^*-\delta}$ *and* v *on* $\partial\Omega_-$. As above, only two cases may occur:

(i′) There exists $\delta_2 \in (0, \delta_0)$ such that $w_{\tau^*-\delta_2} > v$ on $\partial\Omega_-$.

(ii′) For n large enough, there exists $(x_n, y_n) \in \partial\Omega_-$ such that

$$w_{\tau^*-1/n}(x_n, y_n) \leq v(x_n, y_n).$$

If case (i′) occurs, then, for any $0 \leq \delta \leq \delta_2$, we have $w_{\tau^*-\delta} > v$ on $\partial\Omega_-$. Since $f \equiv 0$ on $[0, \theta]$ and $v \leq \theta$ in Ω_-, with the same method as in the proof of Lemma 3.9, we would actually find that $w_{\tau^*-\delta} \geq v$ in Ω_- for all $0 \leq \delta \leq \delta_2$.

If case (ii′) occurs, we can argue word by word as in case (ii) above. That leads to the conclusion of Lemma 3.10.

Completion of the proof of Lemma 3.10. To complete the proof, the only thing left to consider is the case where both (i) and (i′) occur. Set $\delta_3 = \min(\delta_1, \delta_2)$. Thus

$$\forall \delta \in [0, \delta_3], \quad w_{\tau^*-\delta} \geq v \ \text{ in } \Omega_+ \cup \Omega_-. \tag{3.22}$$

From the definition of τ^*, for any $n \geq 1$, there exists a point (x_n, y_n) such that

$$w_{\tau^*-1/n}(x_n, y_n) < v(x_n, y_n).$$

By (3.22), the points (x_n, y_n) are in $\mathcal{B}$ for n large enough. Consequently, up to extraction of a subsequence, one of the following situations occurs:

(i,i′)(a) $X_n \to -\infty$, $Y_n \to \overline{Y} \in [0, A]$.

(i,i′)(b) $x_n \to +\infty$, $y_n \to \overline{y} \in [0, A \sin\alpha]$. The latter can be treated in the same way as the case (ii)(b) above: it is ruled out by the strong maximum principle.

Hence, only case (i,i′)(a) may occur and, as in the case (ii)(a), we get the conclusion of Lemma 3.10. □

Fourth step: Proving the planar behavior of u far away from the axis of symmetry. We are going to use here the (X, Y) coordinates. Fix a point $(X, Y) \in \mathbb{R}^2$. With the notation of Lemma 3.10, we have $X \geq X_n$ for n large enough. Since v is nondecreasing in the direction X, it follows that

$$v(X, Y) \geq v(X_n, Y) = v_n(0, Y - Y_n)$$

for n large enough. Since $Y_n \to \overline{Y}$ and since v has bounded derivatives, we conclude from Lemma 3.10 that

$$v(X_n, Y) \to U(\tau^* + Y) \text{ as } n \to \infty,$$

whence

$$v(X, Y) \geq U(\tau^* + Y).$$

On the other hand, from the definition of τ^*, we have

$$v(X, Y) \leq U(\tau^* + Y - \phi(X)).$$

By summarizing the previous results, it follows that

$$U(\tau^* + Y) \leq v(X, Y) \leq U(\tau^* + Y - \phi(X)) \text{ in } \mathbb{R}^2. \tag{3.23}$$

Now, for any $X_0 \geq 0$, consider the function

$$w^{X_0}(x, y) = U(Y - \phi(X - X_0)).$$

We could compare the functions w^{X_0} and v by arguing in the same way as above. First, the function w^{X_0} satisfies (3.13). Second, instead of (3.14), it is easy to check that

$$\forall \tau \geq X_0 \cot\alpha, \quad w_\tau^{X_0} := U(\tau + Y - \phi(X - X_0)) \geq v \text{ in } E_0.$$

Furthermore, we have $Y - \phi(X - X_0) \geq -X_0 \cot\alpha$ on ∂E_0. Hence, there exists a real $\tau_1' \geq 0$ that we can choose greater than $X_0 \cot\alpha$ such that

$$\forall \tau \geq \tau_1', \quad w_\tau^{X_0} \geq 1 - \varepsilon \text{ on } \partial E_0$$

with the same ε as in (3.16). As in Lemma 3.9, it follows that

$$\forall \tau \geq \tau_1', \quad w_\tau^{X_0} \geq v \text{ in } \mathbb{R}^2.$$

Lemma 3.10 can be applied to the function w^{X_0}. As for (3.23), we get the existence of a real $\tilde{\tau}^*$ such that

$$U(\tilde{\tau}^* + Y) \leq v(X, Y) \leq U(\tilde{\tau}^* + Y - \phi(X - X_0)) \text{ in } \mathbb{R}^2. \tag{3.24}$$

By taking the limit $X \to -\infty$ in (3.23) and (3.24) and by using the monotonicity of U, we conclude that $\widetilde{\tau^*} = \tau^*$.

As a consequence, for all $X_0 \geq 0$, we have

$$U(\tau^* + Y) \leq v(X, Y) \leq U(\tau^* + Y - \phi(X - X_0)) \ \text{ in } \mathbb{R}^2.$$

We pass to the limit $X_0 \to +\infty$ and obtain

$$U(\tau^* + Y) \leq v(X, Y) \leq U(\tau^* + Y) \ \text{ in } \mathbb{R}^2.$$

Since $v(0, 0) = U(0) = \theta$, it follows that $\tau^* = 0$. In other words, the function v is actually nothing but the planar function $U(Y)$. Last, the function v, which is the limit of a subsequence of the functions $v_n(x, y) = u(x + x_n, y + y_n)$, does not depend on the sequence $x_n \to -\infty$. We conclude that the whole sequence (u_n) approaches the function $U(Y)$.

So far, we have proved that, for any $x \in \mathbb{R}$, there existed a unique real $y = \varphi_\theta(x)$ such that $u(x, y) = \theta$. Furthermore, for any sequence $x_n \to -\infty$, the functions $u_n(x, y) = u(x + x_n, y + \varphi_\theta(x_n))$ approach the planar function $U(Y) = U(y \sin\alpha - x \cos\alpha)$.

Let $\lambda \in (0, 1)$. We shall now prove that the level set $\{(x, y),\ u(x, y) = \lambda\}$ is a curve $\{y = \varphi_\lambda(x),\ x \in \mathbb{R}\}$.

First of all, the function u is increasing with respect to y. For each $x \in \mathbb{R}$, set $\psi(x) = \lim_{y \to +\infty} u(x, y)$. In the set $\Omega = \mathbb{R} \times (0, 1)$, let us define the functions

$$\tilde{u}_n(x, y) = u(x, y + n) \ \text{ in } \Omega.$$

They still satisfy (3.1). From the standard elliptic estimates, those functions $\tilde{u}_n$ approach, up to extraction of some subsequence, a function u_∞ that is a solution of

$$\Delta u_\infty - c\partial_y u_\infty + f(u_\infty) = 0 \ \text{ in } \Omega.$$

But this function $v_\infty(x, y)$ is actually identically equal to the function $\psi(x)$. Hence, ψ fulfills

$$\psi'' + f(\psi) = 0 \ \text{ in } \mathbb{R}.$$

On the other hand, for any $y \in \mathbb{R}$, the function $x \mapsto u(x, y)$ is symmetric, nonincreasing in x for $x \leq 0$, and nondecreasing for $x \geq 0$. The same property holds well for the limit function ψ. Thus, 0 is a minimum point of ψ; whence $\psi''(0) \geq 0$. Furthermore, $\psi''(0) = -f(\psi(0)) \leq 0$. Hence, $\psi''(0) = f(\psi(0)) = 0$. In other words, $\psi(0)$ is a zero of the function f. Since $\psi(0) > u(0, 0) = \theta$ and since f is positive on $(\theta, 1)$, we conclude that $\psi(0) = 1$ and finally that $\psi \equiv 1$.

Hence, for any $x \in \mathbb{R}$, $u(x, y) \to 1$ as $y \to +\infty$. Furthermore, $u(x, y) \to 0$ as $y \to -\infty$ from (3.5) applied in $z_0 = (0, 0)$. Since u is continuous and increasing in y, we conclude that there exists a unique $y = \varphi_\lambda(x)$ such that $u(x, \varphi_\lambda(x)) = \lambda$.

Let (x_n) be a sequence such that $x_n \to -\infty$ as $n \to \infty$ and let K be the compact set

$$K = \{(X, Y) \in \mathbb{R}^2,\ |X| \leq 2 \cot\alpha\ |U^{-1}(\lambda)|,\ |Y| \leq 2|U^{-1}(\lambda)|\}.$$

We know that the functions $u_n(x, y) = u(x + x_n, y + \varphi_\theta(x_n))$ approach the function $U(Y) = U(y \sin\alpha - x \cos\alpha)$ uniformly in K. For any $\varepsilon > 0$, there exists an integer n_0 such that if $n \geq n_0$, then

$$u_n(0, (1/\sin\alpha)\ U^{-1}(\lambda) - \varepsilon) < \lambda \ \text{ and } \ u_n(0, (1/\sin\alpha)\ U^{-1}(\lambda) + \varepsilon) > \lambda.$$

Hence, for $n \geq n_0$, one has

$$\varphi_\theta(x_n) + (1/\sin\alpha)\ U^{-1}(\lambda) - \varepsilon \leq \varphi_\lambda(x_n) \leq \varphi_\theta(x_n) + (1/\sin\alpha)\ U^{-1}(\lambda) + \varepsilon.$$

It then follows that

$$\varphi_\lambda(x_n) - \varphi_\theta(x_n) \to (1/\sin\alpha)\ U^{-1}(\lambda) \text{ as } n \to \infty.$$

Since this limit does not depend on the sequence $x_n \to -\infty$, we conclude that, for any $\lambda, \lambda' \in (0,1)$,

$$\varphi_\lambda(x) - \varphi_{\lambda'}(x) \to (1/\sin\alpha)\ (U^{-1}(\lambda) - U^{-1}(\lambda')) \text{ as } x \to -\infty.$$

The same limit also holds as $x \to +\infty$ by symmetry.

In particular, that implies that the functions $\tilde{u}_n(x,y) = u(x + x_n, y + \varphi_\lambda(x_n))$ approach the function $U(Y + U^{-1}(\lambda))$ in $W^{2,p}_{loc}(\mathbb{R}^2)$.

3.4. Asymptotic directions for the level sets of *u*. Let $\vec{k}$ be a vector in the open cone $\mathcal{C}(\vec{e_2}, \pi - \alpha)$. We are going to prove that the function u fulfills the limiting condition (1.3), namely, that $u(\lambda\vec{k}) \to 1$ as $\lambda \to +\infty$. By symmetry with respect to x and since $u(0,y) \to 1$ as $y \to +\infty$, it is enough to treat the case of a vector $\vec{k}$ such that $\vec{k} \cdot \vec{e}_1 < 0$. We can write $\vec{k} = (-\sin\beta, -\cos\beta)$ with $\alpha < \beta < \pi$ (β is the angle between $\vec{k}$ and $-\vec{e}_2$ if one goes clockwise).

Let $0 < \varepsilon < 1$. We shall show that, for λ large enough, we have

$$u(\lambda\vec{k}) \geq 1 - \varepsilon.$$

Consider the compact $K = [-1,1] \times [-2\cot\alpha, 2\cot\alpha]$ and the functions

$$u_n(x,y) = u(x - n, y + \varphi_{1-\varepsilon/2}(-n)).$$

From the previous sections, these functions u_n converge uniformly in K to the function $U(y\sin\alpha - x\cos\alpha + U^{-1}(1 - \varepsilon/2))$.

Let S be the segment between the points $(0,0)$ and $(-1, -\cot\alpha)$. The functions u_n converge uniformly to $1 - \varepsilon/2$ on S. Since u is increasing in y, we deduce that there exists n_0 large enough such that

$$(3.25)\ \forall n \geq n_0, \quad \forall x \in [-n-1, -n], \quad \varphi_{1-\varepsilon}(x) \leq \varphi_{1-\varepsilon/2}(-n) + \cot\alpha\ (x + n).$$

Similarly, since $\alpha < \beta < \pi$ and since U is increasing, the sequence $(u_n(-1, -\cot((\alpha + \beta)/2)))$ approaches $1 - \eta$, as $n \to \infty$, with $0 < \eta < \varepsilon/2$. Hence, there exists $n_0' \geq n_0$ such that

$$\forall n \geq n_0', \quad \varphi_{1-\varepsilon/2}(-n-1) \leq \varphi_{1-\varepsilon/2}(-n) - \cot((\alpha+\beta)/2).$$

With an immediate induction, we get that

$$(3.26) \qquad \forall n \geq n_0', \quad \varphi_{1-\varepsilon/2}(-n) \leq \varphi_{1-\varepsilon/2}(-n_0') - \cot((\alpha+\beta)/2)(n - n_0').$$

Putting together (3.25) and (3.26), we have, for all $n \geq n_0'$ and for all $x \in [-n-1, -n]$,

$$\varphi_{1-\varepsilon}(x) \leq \varphi_{1-\varepsilon/2}(-n_0') + \cot\alpha\ (x + n) - \cot((\alpha+\beta)/2)\ (n - n_0').$$

Since $\cot\alpha \geq \cot((\alpha+\beta)/2)$ and since $x+n \leq 0$ in the previous inequality, we get

$$\forall x \leq -n_0', \quad \varphi_{1-\varepsilon}(x) \leq \varphi_{1-\varepsilon/2}(-n_0') + \cot((\alpha+\beta)/2)\ (x+n_0').$$

By putting $x = -\lambda\sin\beta$ in the last inequality, and since $\beta > \alpha$, we conclude that, for λ large enough,

$$\varphi_{1-\varepsilon}(-\lambda\sin\beta) \leq -\lambda\cos\beta.$$

Remember that $\vec{k} = (-\sin\beta, -\cos\beta)$ and that u is increasing with respect to y. It follows that $u(\lambda\vec{k}) \geq 1-\varepsilon$ for λ large enough. That implies the required formula (1.3).

Since (1.3) is true for any $\vec{k} \in \mathcal{C}(\vec{e}_2, \pi-\alpha)$ and since u is increasing with respect to y, the stronger limit (1.9) also holds.

Furthermore, for any $\rho \in \mathcal{C}(-\vec{e}_2, \alpha)$, we already know that u is nonincreasing in the direction ρ. Hence, for any $\tau > 0$, the function $z = u((x,y)+\tau\rho) - u(x,y)$ is nonpositive and it satisfies a linear elliptic equation of the type $\Delta z - c\partial_z + c(x,y)z = 0$ in $\mathbb{R}^2$ where $c(x,y)$ is a bounded function. Since $u(\lambda\rho) \to 0$ (resp., 1) as $\lambda \to +\infty$ (resp., $\lambda \to -\infty$), the function z cannot be identically 0. The strong maximum principle implies then that $z > 0$ in $\mathbb{R}^2$. In other words, the function u is decreasing in the direction ρ.

Last, the limiting conditions (1.2) and (1.3) imply that each level set $\{y = \varphi_\lambda(x),\ x \in \mathbb{R}\} = \{u = \lambda\}$ of the function u has two asymptotic directions that are directed by the vectors $(\pm\sin\alpha, -\cos\alpha)$.

4. Uniqueness of the speed *c*. In sections 2 and 3, we have proved the existence of a solution (c,u) of (1.1)–(1.3), (1.8)–(1.9) with the speed $c = c_0/\sin\alpha$ for any angle $\alpha \in (0, \pi/2]$.

Choose an angle $\alpha \in (0,\pi/2]$ and let (c,u) be a solution of (1.1)–(1.3), (1.8)–(1.9). First of all, since f is extended by 0 outside $[0,1]$, the strong maximum principle implies that $0 < u < 1$ in $\mathbb{R}^2$. We shall now prove the equality $c = c_0/\sin\alpha$. We divide the proof into three main steps.

(1) Let us consider the case where $0 < \alpha < \pi/2$ and let us suppose that $c < c_0/\sin\alpha$. For $\varepsilon > 0$ small enough, let f_ε be the function defined in $[-\varepsilon, 1-\varepsilon]$ by

$$f_\varepsilon(s) = \begin{cases} f(s) & \text{on } [-\varepsilon, 1-2\varepsilon], \\ \min\ (f(s), (1-\varepsilon-s)/\varepsilon\ f(1-2\varepsilon)) & \text{on } [1-2\varepsilon, 1-\varepsilon]. \end{cases}$$

Furthermore, we extend the functions f_ε by 0 outside $[-\varepsilon, 1-\varepsilon]$. For $\varepsilon > 0$ small enough, f_ε is Lipschitz continuous in $[-\varepsilon, 1-\varepsilon]$, $(f_\varepsilon)'_-(1-\varepsilon) := \lim_{t\to 1-\varepsilon,\ t<1-\varepsilon} \frac{f_\varepsilon(t)}{t-1+\varepsilon}$ exists and is negative, and f_ε fulfills (1.4) on $[-\varepsilon, 1-\varepsilon]$ with the ignition temperature θ. Moreover, we have $f_\varepsilon \leq f$ and the functions f_ε approach f uniformly in $[0,1]$ as $\varepsilon \to 0$. From the results in [2], [9], [15], [24], there exists a unique couple $(c_\varepsilon, u_\varepsilon)$ satisfying

$$\text{(4.1)} \qquad \begin{cases} u_\varepsilon'' - c_\varepsilon u_\varepsilon' + f_\varepsilon(u_\varepsilon) &= 0 \quad \text{in } \mathbb{R}, \\ u_\varepsilon(-\infty) = -\varepsilon,\ u_\varepsilon(0) = \theta,\ u_\varepsilon(+\infty) &= 1-\varepsilon. \end{cases}$$

Furthermore, we have $c_\varepsilon \leq c_0$ and $c_\varepsilon \to c_0$ as $\varepsilon \to 0$ [9].

Since $c < c_0/\sin\alpha$ and $0 < \alpha < \pi/2$, there exist a real $\varepsilon > 0$ small enough and an angle α' such that $0 < \alpha < \alpha' < \pi/2$ and $c < c_\varepsilon/\sin\alpha' < c_0/\sin\alpha$. Set

$$v(x,y) = u_\varepsilon(y\sin\alpha' - x\cos\alpha').$$

Let us first check that v is a subsolution of (1.1). Indeed,

$$\begin{aligned}\Delta v - c\partial_y v + f(v) &= u_\varepsilon'' - c\ \sin\alpha'\ u_\varepsilon' + f(u_\varepsilon)\\ &= (c_\varepsilon - c\sin\alpha')u_\varepsilon' + f(u_\varepsilon) - f_\varepsilon(u_\varepsilon) > 0 \text{ in } \mathbb{R}^2\end{aligned} \tag{4.2}$$

since $c_\varepsilon > c\ \sin\alpha'$, $u_\varepsilon' > 0$, and $f \geq f_\varepsilon$.

We now claim that there exists $\tau \geq 0$ such that

$$v(x, y-\tau) < u(x,y) \text{ in } \mathbb{R}^2. \tag{4.3}$$

If not, then for any $n \in \mathbb{N}$, there exists a point $(x_n, y_n) \in \mathbb{R}^2$ such that

$$v(x_n, y_n - n) = u_\varepsilon(\sin\alpha'\ (y_n - n) - \cos\alpha'\ x_n)\ \geq\ u(x_n, y_n). \tag{4.4}$$

The points (x_n, y_n) are not bounded; otherwise the left-hand side of (4.4) approaches $-\varepsilon$, whereas the right-hand side is nonnegative. Write $(x_n, y_n) = \lambda_n(\sin\varphi_n, -\cos\varphi_n)$ with $-\pi < \varphi_n \leq \pi$: φ_n is the angle between (x_n, y_n) and the vector $-\vec{e}_2$ if one goes counterclockwise. We have $\lambda_n \to +\infty$. We can assume, up to extraction, that the sequence (φ_n) approaches $\varphi \in [-\pi, \pi]$ as $n \to +\infty$.

If $-\alpha' < \varphi < \pi - \alpha'$, then

$$v(x_n, y_n - n) = u_\varepsilon(-\lambda_n \sin(\alpha' + \varphi_n) - n\sin\alpha') \to -\varepsilon \text{ as } n \to \infty.$$

This is ruled out by (4.4) since $u > 0$.

In the other case, one has $-\pi \leq \varphi \leq -\alpha'$ or $\pi - \alpha' \leq \varphi \leq \pi$. In particular, $\varphi \in [-\pi, -\alpha) \cup (\alpha, \pi]$. The limiting condition (1.9) implies that $u(x_n, y_n) \to 1$ as $n \to \infty$. This contradicts (4.4) because $u_\varepsilon \leq 1 - \varepsilon$.

As a consequence, (4.3) is true. Next, decrease τ and define

$$\tau^* = \inf\{\tau \in \mathbb{R},\ v(x, y-\tau) < u(x,y) \text{ in } \mathbb{R}^2\}.$$

This real τ^* is finite because there are some points (x, y) where $u(x,y) < 1 - \varepsilon$ and $v(x, y-\tau) \to 1 - \varepsilon$ as $\tau \to -\infty$. For each $n \in \mathbb{N}^*$, there exists a point (x^n, y^n) such that

$$v(x^n, y^n - \tau^* + 1/n) = u_\varepsilon(\sin\alpha'\ (y^n - \tau^* + 1/n) - \cos\alpha'\ x^n) \geq u(x^n, y^n).$$

With the same arguments as above, we claim that the points (x^n, y^n) are bounded. Hence there exists a point $(\overline{x}, \overline{y}) \in \mathbb{R}^2$ such that $v(\overline{x}, \overline{y} - \tau^*) \geq u(\overline{x}, \overline{y})$. Moreover, owing to the definition of τ^*, we have $v(x, y-\tau^*) \leq u(x,y)$ in $\mathbb{R}^2$. The function $z(x,y) = v(x, y-\tau^*) - u(x,y)$ is nonpositive and reaches 0 somewhere in $\mathbb{R}^2$. Furthermore, from (1.1) and (4.2), it satisfies $\Delta z - c\partial_y z + f(v(x, y-\tau^*)) - f(u) \geq 0$ in $\mathbb{R}^2$. This implies that

$$\Delta z - c\partial_y z + c(x,y) z \geq 0$$

for a bounded function $c(x,y)$. The strong maximum principle yields that $z \equiv 0$ in $\mathbb{R}^2$; i.e., $v(x, y-\tau^*) = u_\varepsilon(\sin\alpha'\ (y - \tau^*) - \cos\alpha'\ x) \equiv u(x,y)$ in $\mathbb{R}^2$. This is impossible because $u_\varepsilon \leq 1 - \varepsilon$ and $\sup_{\mathbb{R}^2} u = 1$.

Eventually, that shows that if $0 < \alpha < \pi/2$, then $c \geq c_0/\ \sin\alpha$.

(2) In this part, we deal with the case $\alpha = \pi/2$, which has not been treated in part 1. Indeed, the sliding method used in part 1 no longer works for the limiting case $\alpha = \pi/2$.

Suppose that $c < c_0$. With the same notation as in part 1, there exists a real $\varepsilon > 0$, small enough and fixed, such that $c < c_\varepsilon$, where $(c_\varepsilon, u_\varepsilon)$ is the solution of (4.1). For some reals η, $\kappa > 0$ that will be chosen later, consider the function $v(x, y) = u_\varepsilon(y - \varphi(x))$, where $\varphi(x) = \sqrt{\eta^2 x^2 + \kappa^2}$.

Let us check that this function v is a subsolution of (1.1) if $\eta > 0$ and $\kappa > 0$ are suitably chosen. We have

$$\begin{aligned}\Delta v - c\partial_y v + f(v) &= (1 + \varphi'(x)^2)u_\varepsilon'' - \varphi''(x)u_\varepsilon' - cu_\varepsilon' + f(u_\varepsilon)\\ &= \varphi'(x)^2 u_\varepsilon'' + (c_\varepsilon - c - \varphi''(x))u_\varepsilon' + f(u_\varepsilon) - f_\varepsilon(u_\varepsilon).\end{aligned}$$

On the one hand, we have $f \geq f_\varepsilon$. On the other hand, since u_ε fulfills (4.1), it is well known that u_ε admits the following asymptotic behavior as $x_1 \to \pm\infty$: $u_\varepsilon(x_1) = -\varepsilon + (\theta + \varepsilon)e^{c_\varepsilon x_1}$ if $x_1 \leq 0$ and $u_\varepsilon(x_1) = 1 - \varepsilon - \alpha e^{\lambda' x_1} + o(e^{\lambda x_1})$, $u_\varepsilon'(x_1) = -\alpha\lambda e^{\lambda' x_1} + o(e^{\lambda x_1})$ as $x_1 \to +\infty$, where $\lambda = \frac{c_\varepsilon - \sqrt{c_\varepsilon^2 - 4(f_\varepsilon)'_-(1-\varepsilon)}}{2} < 0$. Furthermore, we have $u_\varepsilon'' = c_\varepsilon u_\varepsilon' - f_\varepsilon(u_\varepsilon)$ and $u_\varepsilon' > 0$ in $\mathbb{R}$. Finally, there exists a constant $C > 0$ such that $|u_\varepsilon''| \leq Cu_\varepsilon'$ in $\mathbb{R}$. Remember now that $c_\varepsilon > c$. In order to have $\Delta v - c\partial_y v + f(v) \geq 0$ in $\mathbb{R}^2$, it is then sufficient to choose the function ϕ such that $|\varphi'^2|$ and $|\varphi''|$ are small enough. We have $|\varphi'^2| \leq \eta^2$ and $|\varphi''| \leq \eta^2/\kappa$. Hence, we can choose $\eta > 0$ and $\kappa > 0$ such that

$$\Delta v - c\partial_y v + f(v) \geq 0 \quad \text{in } \mathbb{R}^2.$$

To sum up, the function v is a subsolution of (1.1) and each of its level sets has two asymptotes directed by the vectors $(\pm 1, \arctan \eta)$.

We can now argue as in part 1: formula (4.3) is still true if τ is large enough. As in part 1, we can decrease τ, we can define τ^*, and we get a contradiction thanks to the maximum principle.

This eventually proves that if $\alpha = \pi/2$, then $c \geq c_0$.

(3) Choose now any angle $\alpha \in (0, \pi/2]$. We still have to prove that $c \leq c_0/\sin\alpha$. Suppose on the contrary that $c > c_0/\ \sin\alpha$. Let us consider some functions f^ε on $[\varepsilon, 1+\varepsilon]$ such that $f^\varepsilon = f$ on $[\varepsilon, 1-\varepsilon]$, $f^\varepsilon > 0$ on $(\theta, 1+\varepsilon)$, $f^\varepsilon(1+\varepsilon) = 0$, $(f^\varepsilon)'(1+\varepsilon)$ exists and is negative, $f^\varepsilon \geq f$ and $\|f^\varepsilon - f\|_\infty \to 0$ as $\varepsilon \to 0$. In particular, the function f^ε is of the ignition temperature type on the interval $[\varepsilon, 1+\varepsilon]$. For each $\varepsilon > 0$ small enough, there exists a unique couple $(c^\varepsilon, u^\varepsilon)$ fulfilling

$$\left\{\begin{aligned} u^{\varepsilon\prime\prime} - c^\varepsilon u^{\varepsilon\prime} + f^\varepsilon(u^\varepsilon) &= 0 \quad \text{in } \mathbb{R},\\ u^\varepsilon(-\infty) = \varepsilon,\ u^\varepsilon(0) = \theta,\ u^\varepsilon(+\infty) &= 1+\varepsilon.\end{aligned}\right.$$

Furthermore, $c^\varepsilon > c_0$ and $c^\varepsilon \to c_0$ as $\varepsilon \to 0$ (see [9]).

Choose α' and $\varepsilon > 0$ such that $0 < \alpha' < \alpha \leq \pi/2$ and $c > c^\varepsilon/\sin\alpha' > c_0/\ \sin\alpha$. From Theorem 1.1 applied to the function f^ε, there exists a solution $v(x, y)$ of

$$\left\{\begin{array}{ll} & \Delta v - c^\varepsilon/\sin\alpha'\ \partial_y v + f^\varepsilon(v) = 0 \ \text{ in } \mathbb{R}^2,\\ v(\lambda\vec{k}') \to \varepsilon & \text{as } \lambda \to +\infty \text{ and } \vec{k}' \to \vec{k} \in \mathcal{C}(-\vec{e}_2, \alpha'),\\ v(\lambda\vec{k}') \to 1+\varepsilon & \text{as } \lambda \to +\infty \text{ and } \vec{k}' \to \vec{k} \in \mathcal{C}(\vec{e}_2, \pi - \alpha').\end{array}\right.$$

Moreover, $\partial_y v \geq 0$. The function v is a supersolution of (1.1) in the sense that

$$\Delta v - c\partial_y v + f(v) = (c^\varepsilon/\sin\alpha' - c)\partial_y v + f(v) - f^\varepsilon(v) \leq 0 \ \text{ in } \mathbb{R}^2$$

since $c > c^\varepsilon/\sin\alpha'$, $\partial_y v \geq 0$, and $f \leq f^\varepsilon$.

We now claim that there exists $\tau \geq 0$ such that

$$v(x, y + \tau) > u(x, y) \text{ in } \mathbb{R}^2.$$

Otherwise, for each $n \in \mathbb{N}$, there exists a point $(x^n, y^n) \in \mathbb{R}^2$ such that $v(x^n, y^n + n) \leq u(x^n, y^n)$. As in part 1, by dealing successively with the cases where the sequence (x_n, y_n) is bounded or unbounded, we would get a contradiction.

Now, let us set

$$\tau^* = \inf \{\tau \in \mathbb{R},\ v(x, y + \tau) > u(x, y) \text{ in } \mathbb{R}^2\}.$$

As above, τ^* is finite and $v(x, y + \tau^*) \geq u(x, y)$ in $\mathbb{R}^2$ with equality somewhere. This is ruled out by the strong maximum principle.

Finally, it is always true that $c \leq c_0/\ \sin\alpha$. Together with parts 1 and 2, this inequality completes the proof of Theorem 1.2.

5. Appendix: Proof of Lemma 2.10. In this section, we actually deal with a more general situation than in Lemma 2.10. Let u be a bounded and positive function defined in the set

$$V = \{(x, y) \in \mathbb{R}^2, x > 0, y > 0, \sqrt{x^2 + y^2} < \delta\}$$

for a certain $\delta > 0$. We assume that the function u belongs to $W^{2,p}_{loc}(\overline{V}\backslash\{(0,0)\})$ for all $1 < p < \infty$ and that it is continuous in $\overline{V}$. We also suppose that that function v satisfies the following equations:

$$\left\{ \begin{aligned} \Delta u - c\partial_y u + f(u) &= 0 \quad \text{in } V, \\ u(x, 0) &= 0 \quad \text{for } 0 \leq x \leq \delta, \\ \partial_\tau u(0, y) &= 0 \quad \text{for } 0 < y \leq \delta, \end{aligned} \right. \tag{5.1}$$

where $\tau = (-\sin\alpha, -\cos\alpha)$. The given function f is Lipschitz continuous. Furthermore, $f(0) = 0$ and $f'_+(0) = \lim_{t\to 0,\ t>0} \frac{f(t)-f(0)}{t}$ exists.

Set $O = (0, 0)$. Choose any vector $\rho = (\cos\beta, \sin\beta)$ with $\pi/2 - \alpha < \beta < \pi$. We are going to determine the asymptotic behavior of u and ∇u in the neighborhood of the corner O. That behavior will imply the existence of a neighborhood $\tilde{V}$ of O and of a real $\varepsilon_1 > 0$ such that if $0 < \varepsilon \leq \varepsilon_1$ and if z, $z + \varepsilon\rho \in \tilde{V} \cap \overline{V}$, then $u(z) < u(z + \varepsilon\rho)$.

Before doing that, we briefly mention some papers and results that have been devoted to similar problems in the literature. In many works (see, e.g., Bernardi and Maday [10], Grisvard [19], Maz'ja and Plamenevskii [30]), the *linear* elliptic problem

$$\begin{gathered} Lu = f \text{ in } G, \\ Bu = g \text{ on } \partial G\backslash\{K\} \end{gathered} \tag{5.2}$$

has been investigated under the assumption that G is a subdomain of the plane $\mathbb{R}^2$ and that the boundary ∂G of G is Lipschitz continuous everywhere and smooth except at a corner K, say, $K = O$. Assume that L is an elliptic operator and B is a smooth linear function depending on the traces of u or ∇u on $\partial G\backslash\{K\}$. The function u belongs to some Sobolev spaces with weights but u, or its derivatives, may be singular at the point K. The general result is the following: in a neighborhood of the point $K = O$, the function u can be written as

$$u(r, \theta) = \sum_{k\geq 1} c_k r^{\alpha_k} \sum_{h=0}^{k} (-\ln r)^h \varphi_{k,h}(\theta), \tag{5.3}$$

where (r, θ) is the usual polar coordinate and where the complex numbers α_k have nondecreasing real parts. Thanks to the change of variables $r = e^t$ (see Kondrat'ev [25]), equation (5.2) becomes

$$\tilde{L}u = \tilde{f}$$

in a set containing an infinite strip of the type $(-\infty, \alpha] \times (0, \beta)$. The terms r^{α_k} become $e^{\alpha_k t}$ and the numbers α_k are given in terms of the eigenvalues of an operator L_0 depending on θ and on the principal part of L at the corner K.

In particular, for the Dirichlet problem

$$\begin{aligned} \Delta u = f \quad & \text{in } G = \{r > 0,\ 0 < \theta < \omega\}, \\ u = 0 \quad & \text{on } \partial G \backslash \{K\}, \end{aligned}$$

where $f \in W^{m,p}(G)$, it is known that, in a neighborhood of K, the function u is equal to

$$u(r, \theta) = \sum_{\pi/\omega \leq k\pi/\omega < m+2-2/p} c_k r^{k\pi/\omega} \left\{ \begin{array}{l} \sin(k\pi\theta/\omega) \\ \text{or } (\ln r)\sin(k\pi\theta/\omega) + \theta\cos(k\pi\theta/\omega) \end{array} \right. + \ u_R,$$

where $u_R \in W^{m+2,p}(G)$ (see Geymonat and Grisvard [16], Grisvard [19], [20], or Dauge [13] for a three-dimensional situation).

Let us now come back to the elliptic problem (5.1) that is set in the domain V with the corner O. The boundary conditions on ∂V are of the Dirichlet and oblique-Neumann type. But, unlike the problems mentioned above, we have to deal with a *semilinear* problem. Then, we cannot a priori hope for an infinite asymptotic development of the type (5.3) for u. Nevertheless, we only need to know what u and its derivatives are equivalent to in the neighborhood of O.

In [9], [8], Berestycki and Nirenberg have emphasized the semilinear problem

$$\begin{aligned} Lu + f(x_1, u) = 0,\ u > 0 \quad & \text{in } \Sigma_- = \{(x_1, y),\ x_1 < 0,\ y \in \omega\}, \\ \partial_\nu u = 0 \quad & \text{on } (-\infty, 0) \times \partial\omega, \end{aligned}$$

where ω is a smooth domain with unit outward normal ν. If $u \to 0$ as $x_1 \to -\infty$ and if $|f(x_1, u)| = O(u^{1+\delta})$ as $u \to 0$ for a certain $\delta > 0$, then the nonlinear term $f(x_1, u)$ only makes small perturbations with respect to Δu. The asymptotic behavior of u as $x_1 \to -\infty$ is given in [8], [9].

If we come back to (5.1) and if we make the change of variables $r = e^t$, we can see that u fulfills

$$\Delta u - c\sin\theta \varepsilon^t \partial_t u - c\cos\theta\ e^t \partial_\theta u + e^{2t} f(u) = 0 \text{ in } (-\infty, \ln\delta) \times (0, \pi/2)$$

with Dirichlet and oblique-Neumann boundary conditions:

$$\begin{aligned} u = 0 \quad & \text{on } \{\theta = 0\}, \\ -\cos\alpha\ \partial_t u + \sin\alpha\ \partial_\theta u = 0 \quad & \text{on } \{\theta = \pi/2\}. \end{aligned}$$

To conclude this discussion, the semilinear problem (5.1) with mixed boundary conditions does not seem to have been treated so far in the literature. Hence, for the sake of completeness, we give a detailed proof of Lemma 5.1.

LEMMA 5.1. *Let $\gamma = (2/\pi)\ \alpha$. There exists a real $\lambda > 0$ such that*

$$\left\{ \begin{array}{rl} u - \lambda r^\gamma \sin(\gamma\theta) & = o(r^\gamma) \\ \nabla u - \lambda \nabla(r^\gamma \sin(\gamma\theta)) & = o(r^{\gamma-1}) \end{array} \right. \quad \textit{as } r \overset{>}{\to} 0.$$

Proof of Lemma 2.10. Consider the behavior of u near the corner C_1 of Σ_a and call (r,θ) the polar coordinates with respect to the point C_1. From Lemma 5.1, one has

$$\nabla u \cdot \rho - \lambda \nabla(r^\gamma \sin(\gamma\theta)) \cdot \rho = o(r^{\gamma-1}) \quad \text{as } r \to 0. \tag{5.4}$$

Remember that $\rho = (\cos\beta, \sin\beta)$ with $\pi/2 - \alpha < \beta < \pi$. Thus,

$$\nabla(r^\gamma \sin(\gamma\theta)) \cdot \rho = \gamma r^{\gamma-1} \sin((\gamma-1)\theta + \beta).$$

For any point $z = (r,\theta) \in V$, we have

$$0 < \alpha - \pi/2 + \beta \leq (\gamma-1)\theta + \beta \leq \beta < \pi.$$

As a consequence, there exists a real $\eta > 0$ such that

$$r^{-(\gamma-1)}\ \nabla(r^\gamma \sin(\gamma\theta)) \cdot \rho \geq \eta > 0.$$

From (5.4), it follows then that $\partial_\rho u > 0$ in a neighborhood V_1 of C_1. As far as the behavior of the function u near the corner C_1 of Σ_a is concerned, Lemma 2.10 is then a consequence of the finite increment theorem.

The other corner C_3 can be treated similarly. Indeed, after setting the origin in C_3 and making the change of variables $y \to -y$, $\tilde{u}(x,y) = u(x,-y)$, we find that

$$\left\{ \begin{array}{ll} (1-\tilde{u}) - \lambda r^\gamma \sin(\gamma\theta) & = o(r^\gamma) \\ -\nabla\tilde{u} - \lambda\nabla(r^\gamma \sin(\gamma\theta)) & = o(r^{\gamma-1}) \end{array} \right. \quad \text{as } r \overset{>}{\to} 0,$$

where $\gamma = (2/\pi)\ (\pi - \alpha)$ and where λ is a positive real. The same calculations as above yield that, for any $\rho = (\cos\beta, \sin\beta)$ with $\pi/2 - \alpha < \beta < \pi$, the function u is such that $\partial_\rho u > 0$ in a neighborhood V_3 of C_3. Notice that, unlike the situation around the point C_1, the function $\partial_\rho u$ is bounded near C_3 since $\gamma \geq 1$. □

Proof of Lemma 5.1. Remember first that $V = \{0 < r < \delta,\ 0 < \theta < \pi/2\}$. We choose to work with the (r,θ) coordinates. Notice that everything works similarly with the coordinates (t,θ), where $r = e^t$. The following proof, similar to the one in [8], is divided into six main steps for the sake of clarity.

Step 1. Set $\gamma = (2/\pi)\ \alpha$; notice that $\gamma \in (0,1]$. Let v be the function

$$v(r,\theta) = r^\gamma \sin(\gamma\theta) \quad \text{for } (r,\theta) \in (0,\delta] \times [0,\pi/2]$$

and $v(O) = 0$. It is easy to check that

$$\left\{ \begin{array}{rl} \Delta v = 0 & \text{in } V, \\ \partial_\tau v(0,y) = 0 & \text{if } 0 < y < \delta, \end{array} \right.$$

where $\tau = (-\sin\alpha, -\cos\alpha)$. Moreover, $v(x,0) = 0$ for all $0 \leq x \leq \delta$ and $v(x,y) > 0$ if $y > 0$.

Step 2. We now want to construct two sub- and supersolutions $\underline{v}$ and $\overline{v}$ such that

$$\left\{ \begin{array}{rl} \Delta\underline{v} - c\partial_y \underline{v} + f(\underline{v}) \geq 0 & \text{in } V_0, \\ \underline{v}(x,0) \leq 0 & \text{if } 0 \leq x < \delta_0, \\ \partial_\tau \underline{v}(0,y) < 0 & \text{if } 0 < y < \delta_0, \end{array} \right. \tag{5.5}$$

$$\text{(5.6)}\qquad \begin{cases} \Delta\overline{v} - c\partial_y\overline{v} + f(\overline{v}) \le 0 & \text{in } V_0, \\ \overline{v}(x,0) \ge 0 & \text{if } 0 \le x < \delta_0, \\ \partial_\tau\overline{v}(0,y) > 0 & \text{if } 0 < y < \delta_0, \end{cases}$$

in a small enough neighborhood V_0 of O of the type $V_0 = V \cap B(0,\delta_0)$, where the real $\delta_0 \in (0,\delta]$ will be chosen later.

Consider the functions

$$\begin{cases} \underline{g}(\theta) & = 1 - \cos(\underline{\beta}\theta) + \underline{A}\sin(\underline{\beta}\theta), \\ \overline{g}(\theta) & = -1 + \cos(\overline{\beta}\theta) + \overline{A}\sin(\overline{\beta}\theta), \end{cases}$$

and

$$\begin{cases} \underline{v} & = r^\gamma\sin(\gamma\theta) + r^{\underline{\beta}}\underline{g}(\theta), \\ \overline{v} & = r^\gamma\sin(\gamma\theta) + r^{\overline{\beta}}\overline{g}(\theta), \end{cases}$$

where $\underline{\beta}$ and $\overline{\beta}$ are two fixed reals, different from 1 and such that $\gamma < \underline{\beta}, \overline{\beta} < \gamma + 1$. The reals $\underline{A}$ and $\overline{A}$ will be chosen later. A straightforward computation gives

$$\begin{aligned} L\underline{v} := \;& \Delta\underline{v} - c\partial_y\underline{v} + f(\underline{v}) \\ = \;& \underline{\beta}^2 r^{\underline{\beta}-2} - c\gamma r^{\gamma-1}\cos((\gamma-1)\theta) \\ & -c\underline{\beta}r^{\underline{\beta}-1}[\sin\theta + \sin((\underline{\beta}-1)\theta) + \underline{A}\cos((\underline{\beta}-1)\theta)] + f(\underline{v}). \end{aligned}$$

Since $\underline{\beta} < \gamma + 1$ and $|f(t)| \le M|t|$ for all t (with $M = \|f\|_{Lip} = \sup_{x,y\in[0,1],\ x\neq y} \frac{|f(x)-f(y)|}{|x-y|}$), it follows that there exists a real $\delta_1 \in (0,\delta]$ that depends only on α, $\underline{\beta}$, M, and $\underline{A}$ such that $L(\kappa\underline{v}) > 0$ in $V \cap B(O,\delta_1)$ for any $\kappa > 0$. On the other hand,

$$\forall 0 < y < \delta, \quad \partial_\tau\underline{v}(0,y) = \underline{\beta}r^{\underline{\beta}-1}[2\sin(\alpha - \underline{\beta}\pi/4)\sin(\underline{\beta}\pi/4) + \underline{A}\sin(\alpha - \underline{\beta}\pi/2)].$$

Since $(2/\pi)\,\alpha < \underline{\beta} < (2/\pi)\,\alpha + 1$, we can then choose a real $\underline{A}$ large enough, depending on α and $\underline{\beta}$, such that $\partial_\tau\underline{v}(0,y) < 0$ for all $0 < y < \delta_1$. Furthermore, we have $\underline{v}(x,y) = 0$ if $y = 0$ and $0 \le x < \delta_1$. We then conclude that $\underline{v}$ satisfies (5.5) in $V \cap B(O,\delta_1)$.

Similarly, we can prove that there exists a real $\delta_2 \in (0,\delta]$ such that $\overline{v}$ satisfies (5.6) in $V \cap B(O,\delta_2)$. Eventually, by defining $\delta_0 = \min(\delta_1,\delta_2)$, it follows that $\underline{v}$ (resp., $\overline{v}$) satisfies (5.5) (resp., (5.6)) in $V_0 = V \cap B(0,\delta_0)$.

Step 3. Even if it means decreasing $\delta_0 > 0$, we can assume that $\underline{v}$ and $\overline{v}$ are positive in $\overline{V_0} \cap \{y > 0\}$. Indeed, this is possible because $\gamma < \underline{\beta}, \overline{\beta}$, because $\sin(\gamma\theta) > 0$ for $0 < \theta < \pi/2$ and because both functions $\underline{g}(\theta)/\sin(\gamma\theta)$ and $\overline{g}(\theta)/\sin(\gamma\theta)$ are bounded in the interval $\{0 \le \theta \le \pi/2\}$. On the other hand, we define a function

$$\varphi(x,y) = 2e^{\cos\alpha\ +\ \sin\alpha} - e^{1/\delta_0(\cos\alpha\ x\ -\ \sin\alpha\ y\ +\ \sin\alpha\ \delta_0)} \ \text{ in } V_0.$$

We observe that the function φ is positive in $\overline{V_0}$ and $\partial_\tau\varphi(0,y) = 0$ for all $0 < y < \delta_0$. Furthermore, we have

$$\Delta\varphi - c\partial_y\varphi + \|f\|_{Lip}\varphi \le -1/\delta_0^2 + 1/\delta_0\ |c|\sin\alpha\ e^{\cos\alpha+\sin\alpha} + 2\|f\|_{Lip}e^{\cos\alpha+\sin\alpha}.$$

Even if it means decreasing again $\delta_0 > 0$, we may also assume that

$$\Delta\varphi - c\partial_y\varphi + \|f\|_{Lip}\varphi < 0 \ \text{ in}V_0.$$

Since u is positive in V_0 and satisfies (5.1), the maximum principle and the Hopf lemma yield that $u(x,y) > 0$ as soon as $y > 0$ and that $\partial_y u(x,0) > 0$ for all $x > 0$. Similarly, $\partial_y \overline{v}(x,0) > 0$ for all $x > 0$. Finally, there exist two reals ν, $\mu > 0$ such that

$$\forall (x,y) \in V \cap \{x^2+y^2=\delta_0^2\}, \quad \mu \underline{v}(x,y) < u(x,y) < \nu \overline{v}(x,y). \tag{5.7}$$

Let us now show that this last inequality (5.7) is actually true in the whole set V_0. Remember that u solves (5.1) and that $\mu \underline{v}$ satisfies inequality (5.5). Hence, the function $w = u - \mu \underline{v}$ satisfies

$$\tilde{L}w := \Delta w - c\partial_y w + c(x,y)w \le 0 \ \text{ in } V_0,$$

where $c(x,y)$ is a bounded function in V_0 such that $\|c\|_\infty \le \|f\|_{Lip}$. Set $g = w/\varphi$. One has

$$Mg := \Delta g + 2\frac{\nabla \varphi}{\varphi} \cdot \nabla g - c\partial_y g \le -\frac{g}{\varphi}(\Delta\varphi - c\partial_y\varphi + c(x,y)\varphi) = -\frac{g}{\varphi}\tilde{L}\varphi.$$

In view of the properties fulfilled by φ, it follows that

$$\tilde{L}\varphi \le \Delta\varphi - c\partial_y\varphi + \|f\|_{Lip}\varphi < 0 \ \text{ in } V_0.$$

If the set $\Omega_- = \{(x,y) \in V_0,\ g(x,y) < 0\}$ is not empty, we get that $Mg < 0$ in Ω_-. Since g is continuous in $\overline{V_0}$ (the function φ is positive and continuous in the compact set $\overline{V_0}$), let z_0 be a point in $\overline{\Omega_-}$ where g reaches its minimal value. If $z_0 \in V_0$, then $\nabla g(z_0) = 0$ and $\Delta g(z_0) \ge 0$. That is impossible because $Mg(z_0) < 0$. Now, since $w \ge 0$ on $\partial V_0 \cap (\{y=0\} \cup \{x^2+y^2=\delta_0^2\})$, it follows that $z_0 = (0,y_0)$ with $0 < y_0 < \delta_0$. Furthermore, since $\partial_\tau \underline{v}(0,y_0) < 0$, we have $\partial_\tau w(z_0) = \partial_\tau u(z_0) - \mu\partial_\tau \underline{v}(z_0) > 0$ and

$$0 < \partial_\tau w(z_0) = g(z_0)\partial_\tau\varphi(z_0) + \varphi(z_0)\partial_\tau g(z_0).$$

The function φ is such that $\partial_\tau \varphi(z_0) = 0$ and $\varphi(z_0) > 0$. Hence, $\partial_\tau g(z_0) > 0$. The latter is ruled out by the Hopf lemma.

Finally, we have $\Omega_- = \emptyset$, whence $w \ge 0$; i.e., $\mu\underline{v} \le u$ in V_0 and even $\mu\underline{v} < u$ in V_0 from the strong maximum principle. Similarly, we infer that $u < \nu\overline{v}$ in V_0.

So far, we have shown that

$$\mu\underline{v} < u < \nu\overline{v} \ \text{ in } V_0 = \{x > 0,\ y > 0,\ r < \delta_0\}.$$

Step 4. Let us now replace the variables (x,y) with $(\varepsilon x, \varepsilon y)$. Set $W_\varepsilon = \{(x,y) \in \mathbb{R}^2,\ (\varepsilon x, \varepsilon y) \in V_0\}$ and $u_\varepsilon(x,y) = \varepsilon^{-\gamma}u(\varepsilon x, \varepsilon y)$ for $(x,y) \in W_\varepsilon$. From the definitions of $\overline{v}$ and $\underline{v}$, we have

$$\mu\ (v + \varepsilon^{\underline{\beta}-\gamma} r^{\underline{\beta}} \underline{g}(\theta)) < u_\varepsilon(x,y) < \nu\ (v + \varepsilon^{\overline{\beta}-\gamma} r^{\overline{\beta}} \overline{g}(\theta)) \ \text{ in } W_\varepsilon, \tag{5.8}$$

where $r = \sqrt{x^2+y^2}$. Let Π be the positive quadrant

$$\Pi = \{x > 0, y > 0\}.$$

Since $\gamma < \underline{\beta}, \overline{\beta}$, the left and the right sides of the inequality (5.8) uniformly approach μv and νv in any compact set $K \subset \overline{\Pi}$ as $\varepsilon \to 0$.

Furthermore, we have

$$\begin{cases} \Delta u_\varepsilon - \varepsilon c \partial_y u_\varepsilon & = -\varepsilon^{2-\gamma} f(u(\varepsilon x, \varepsilon y)) & \text{in } W_\varepsilon, \\ u_\varepsilon(x,0) & = 0 & \text{for all } 0 \le x < \delta_0/\varepsilon, \\ \partial_\tau u_\varepsilon(0,y) & = 0 & \text{for all } 0 < y < \delta_0/\varepsilon. \end{cases}$$

Since $\gamma < 2$ and $f(u)$ is bounded in $\overline{V_0}$, the right side of the equation fulfilled by u_ε approaches 0 uniformly in any compact set $K \subset \overline{\Pi}$. The functions u_ε are defined in such a compact set K for ε small enough and they are also uniformly bounded in K from (5.8). Moreover, from the standard elliptic estimates up to the boundary, the functions (u_ε) are then bounded in $W^{2,p}(K)$ for any compact set $K \subset \overline{\Pi}\backslash\{O\}$ and for any $1 < p < \infty$. By a diagonal extraction process, it follows that there exists a continuous function u_0 defined in $\overline{\Pi}\backslash\{O\}$ such that, up to extraction of some subsequence, $u_\varepsilon \to u_0$ in $C^{1,\delta}_{loc}(\overline{\Pi}\backslash\{O\})$ for any $\delta \in (0,1)$. The function u_0 fulfills

$$\left\{\begin{aligned} \Delta u_0 &= 0 && \text{in } \Pi,\\ u_0(x,0) &= 0 && \text{for all } x > 0,\\ \partial_\tau u_0(0,y) &= 0 && \text{for all } y > 0.\end{aligned}\right. \tag{5.9}$$

Moreover, $\mu v \le u_0 \le \nu v$ in $\overline{\Pi}\backslash\{O\}$. In particular, the latter implies that the function u_0 can be extended by continuity at the point $O = (0,0)$ by setting $u_0(0,0) = 0$. Hence,

$$\mu v \le u_0 \le \nu v \ \text{ in } \overline{\Pi}.$$

From (5.8), for any $\eta > 0$, there exists $\delta' > 0$ such that $|u_\varepsilon| \le \eta$ in $\{(x,y) \in \overline{\Pi},\ \sqrt{x^2+y^2} \le \delta'\}$. It follows that, up to extraction of some subsequence, the functions u_ε also approach u_0 uniformly in any compact set $K \subset \overline{\Pi}$.

Step 5. We now aim at proving that $u_0 = \lambda v$ for a certain λ such that $\mu \le \lambda \le \nu$. Define $\overline{\mu}$ and $\overline{\nu}$ by $\overline{\mu} = \sup\{\mu,\ \mu v \le u_0 \text{ in } \overline{\Pi}\}$ and $\overline{\nu} = \inf\{\nu,\ u_0 \le \nu v \text{ in } \overline{\Pi}\}$. We have $\overline{\mu} v \le u_0 \le \overline{\nu} v$ in $\overline{\Pi}$ and $\overline{\mu} \le \overline{\nu} \in \mathbb{R}$.

Let us now suppose that $\overline{\mu} < \overline{\nu}$. The strong maximum principle then yields that $\overline{\mu} v < u_0 < \overline{\nu} v$ in Π. For every $R > 0$, let us call $C(R) = \{(x,y) \in \overline{\Pi},\ x^2+y^2 = R^2\}$ and $B(R) = \{(x,y) \in \overline{\Pi},\ x^2+y^2 \le R^2\}$. Choose any $R > 0$. On $C(R)$, we have $v > 0$ and $\overline{\mu} \le u_0/v \le \overline{\nu}$. There exists then a subset $\Gamma \subset C(R)$ such that $|\Gamma|/|C(R)| \ge 1/2$ ($|\Gamma|$ is the length of Γ) and one of the following assertions occurs:

$$\text{(i)}\quad \frac{\overline{\mu}+\overline{\nu}}{2} \le \frac{u_0}{v} \ \text{ on } \Gamma, \quad \text{i.e.,}\ \ u_0 - \overline{\mu} v \ge \frac{\overline{\nu}-\overline{\mu}}{2} v,$$

$$\text{(ii)}\quad \frac{u_0}{v} \le \frac{\overline{\mu}+\overline{\nu}}{2} \ \text{ on } \Gamma, \quad \text{i.e.,}\ \ \overline{\nu} v - u_0 \ge \frac{\overline{\nu}-\overline{\mu}}{2} v.$$

Suppose that case (i) occurs. Since $u_0 - \overline{\mu} v > 0$ in Π, since both u_0 and v fulfill (5.9), and since (5.9) is invariant by stretching the variables, a straightforward application of the Harnack inequality up to the boundary leads to the existence of a real $\varepsilon > 0$, which does not depend on R, such that

$$u_0 - \overline{\mu} v \ge \varepsilon v \ \text{ on } C(R/2)$$

(see also Berestycki, Caffarelli, and Nirenberg [3] and Caffarelli [12] for related problems). Hence, as in Step 3, we get

$$u_0 - \overline{\mu} v \ge \varepsilon v \ \text{ in } B(R/2).$$

Since (i) or (ii) occurs for each $R > 0$, we may suppose, say, that there is a sequence $R_n \to +\infty$ such that (i) occurs for each R_n. As a consequence, $u_0 - \overline{\mu} v \ge \varepsilon v$ in $B(R_n/2)$, whence

$$u_0 - \overline{\mu} v \ge \varepsilon v \ \text{ in } \overline{\Pi}.$$

That is ruled out by the definition of $\overline{\mu}$.

We conclude that $\overline{\mu} = \overline{\nu} =: \lambda$, that is to say that $u_0 \equiv \lambda v$ in $\overline{\Pi}$.

Step 6. Conclusion: we have to prove that

$$u - \lambda r^{\gamma}\sin(\gamma\theta) = o(r^{\gamma}) \ \text{ as } r \overset{>}{\to} 0, \tag{5.10}$$

$$\nabla u - \lambda\nabla(r^{\gamma}\sin(\gamma\theta)) = o(r^{\gamma-1}) \ \text{ as } r \overset{>}{\to} 0. \tag{5.11}$$

Let K be the compact defined by $K = \{(x,y) \in \overline{\Pi}, 1 \leq \sqrt{x^2+y^2} \leq 2\}$ and let η be any positive number. We know that $u_\varepsilon \to \lambda v$ as $\varepsilon \to 0$, uniformly in K. Hence, there exists a real $\varepsilon_0 \in (0,1)$ such that: $\forall 0 < \varepsilon \leq \varepsilon_0$, $\forall (x,y) \in K$, $|u_\varepsilon - \lambda v| \leq \eta$. Owing to the definitions of the function u_ε and v, we get

$$\forall (x,y) \in K,\ \forall \varepsilon \leq \varepsilon_0, \quad |u(\varepsilon x, \varepsilon y) - \lambda(\varepsilon r)^{\gamma}\sin(\gamma\theta)| \leq \eta\varepsilon^{\gamma} \leq \eta(\varepsilon r)^{\gamma}.$$

In other words, for each $(x,y) \in \overline{\Pi}$ such that $0 < r = \sqrt{x^2+y^2} \leq 2\varepsilon_0$, we have $|u(x,y) - \lambda r^{\gamma}\sin(\gamma\theta)| \leq \eta r^{\gamma}$. Since $\eta > 0$ was arbitrary, we have thus shown the formula (5.10).

Assertion (5.11) can be proved with the same arguments as above. That completes the proof of Lemma 5.1. □

Remark 5.2. Let $\overline{v}$ be defined as in Step 2 by

$$\overline{v} = r^{\gamma}\sin(\gamma\theta) + r^{\overline{\beta}}\overline{g}(\theta),$$

where $\overline{g}(\theta) = -1 + \cos(\overline{\beta}\theta) + \overline{A}\sin(\overline{\beta}\theta)$ and where (r,θ) are the polar coordinates with respect to the corner $C_1 = (-a, -a\cot\gamma)$ of Σ_a. We choose $\overline{A}$ such that (5.6) holds in $V_0 = \{x > 0,\ y > 0,\ 0 < r < \delta_0\}$ for some δ_0 small enough. In particular, for $\varepsilon \in (0,\delta_0)$, we have $\partial_\tau \overline{v} = \nabla\overline{v}\cdot\tau > 0$ at the point $(-a, -a\cot\gamma+\varepsilon)$. Hence, under the notation of Lemma 2.1, one can require that the vector field ρ_ε fulfill $\rho_\varepsilon = \tau$ on $\{-a\}\times(-a\cot\gamma+\varepsilon, -a\cot\gamma+\delta_0)$ and $\rho_\varepsilon\cdot\nabla\overline{v} \geq 0$ on $\partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0)$. For instance, choose a function $\eta(x,y)$ defined on $\partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0)$ such that $0 \leq \eta \leq 1$, $\eta = 1$ on $\{-a\}\times(-a\cot\gamma+\varepsilon, -a\cot\gamma+\delta_0)$, $\eta = 0$ on $\partial\Sigma_{a,\varepsilon}\cap\{x > -a+\varepsilon^2\}$ (for $\varepsilon > 0$ small enough). Next, take $\rho_\varepsilon(x,y) = \eta(x,y)\tau$ on $\partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0)$. Finally, the function $\overline{v}$ fulfills

$$\rho_\varepsilon\cdot\nabla\overline{v} + \sigma_{0,\varepsilon}\overline{v} \geq 0 \ \text{ on } \partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0),$$

whereas the function u_ε fulfills

$$\rho_\varepsilon\cdot\nabla u_\varepsilon + \sigma_{0,\varepsilon}u_\varepsilon = 0 \ \text{ on } \partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0)$$

(remember that $\sigma_{1,\varepsilon} = 0$ on $\partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0)$ for $\varepsilon > 0$ and $\delta_0 > 0$ small enough).

Furthermore, since $\partial_y u_\varepsilon(-a+\delta_0, -a\cot\gamma) \to \partial_y u_c(-a+\delta_0, -a\cot\gamma) < +\infty$ as $\varepsilon \to 0$ and $u_\varepsilon \leq 1$ in $\overline{\Sigma_{a,\varepsilon}}$, there exists then a constant $\nu > 0$ such that, as in Step 3,

$$\forall (x,y) \in \overline{\Sigma_{a,\varepsilon}}\cap\{r = \delta_0\}, \quad u_\varepsilon(x,y) \leq \nu\overline{v}(x,y)$$

for all $\varepsilon > 0$ small enough. Next, we choose the same function φ as in Step 3. In particular, in view of the choice of ρ_ε, we have $\rho_\varepsilon\cdot\nabla\varphi = 0$ and $\rho_\varepsilon\cdot\nu_\varepsilon \geq 0$ on $\partial\Sigma_{a,\varepsilon}\cap B(C_1,\delta_0)$ for $\varepsilon > 0$ small enough (ν_ε is the outward unit normal to $\partial\Sigma_{a,\varepsilon}$). As in Step 3, it follows then that if the function $g = \frac{w}{\varphi} := \frac{\nu\overline{v}-u_\varepsilon}{\varphi}$ reaches a negative minimal value at a point z_0 in $\overline{\Sigma_{a,\varepsilon}} \cap \overline{B(C_1, \delta_0)}$, then $z_0 = (x_0, y_0)$ lies necessarily on $\partial\Sigma_{a,\varepsilon} \cap B(C_1, \delta_0)$. At the point z_0, one has $\rho_\varepsilon \cdot \nabla w + \sigma_{0,\varepsilon} w \geq 0$, whence

$$g(z_0)\ \rho_\varepsilon(z_0) \cdot \nabla\varphi(z_0) + \varphi(z_0)\ \rho_\varepsilon(z_0) \cdot \nabla g(z_0) + \sigma_{0,\varepsilon}(z_0) g(z_0) \varphi(z_0) \geq 0. \tag{5.12}$$

The first term of (5.12) is equal to 0 because $\rho_\varepsilon \cdot \nabla\varphi = 0$. The second and third terms are nonpositive because $\varphi > 0$, $\rho_\varepsilon \cdot \nabla g \leq 0$ (from the Hopf lemma), $g(z_0) < 0$, and $\sigma_{0,\varepsilon} \geq 0$. Furthermore, if $y_0 \geq -a \cot\gamma + \varepsilon$, then $\rho_\varepsilon(z_0) = \tau$ whence $\rho_\varepsilon(z_0) \cdot \nabla g(z_0) < 0$, and if $y_0 \leq -a \cot\gamma + \varepsilon$, then $\sigma_{0,\varepsilon}(z_0) = 1$. Hence, all the three terms of (5.12) are nonpositive and at least one is negative. This is impossible.

We conclude that

$$u_\varepsilon(x, y) \leq \nu \overline{v}(x, y) \quad \text{in } \overline{\Sigma_{a,\varepsilon}} \cap \overline{B(C_1, \delta_0)}$$

for all $\varepsilon > 0$ small enough. This gives the required estimate (2.5) around the point C_1. The other corners C_2, C_3, C_4 can be treated similarly.

The proofs of the estimates (2.8) and (2.10) resort to the same arguments. As far as (2.8) is concerned, the function $\overline{v}$ can be chosen as in Step 2 such that (5.6) is true for each c_n because the reals c_n are bounded. As far as (2.10) is concerned, the function $\overline{v}$ can be chosen as in Step 2 such that (5.6) is true for each f_n because the norms $\|f_n\|_{Lip}$ are bounded.

Acknowledgments. We are grateful to Prof. P. Clavin for suggesting this subject and to Prof. J.-M. Roquejoffre for useful remarks.

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727; 17 (1964), pp. 35–92.

[2] D. G. Aronson and H. F. Weinberger, *Multidimensional nonlinear diffusions arising in population genetics*, Adv. Math., 30 (1978), pp. 33–76.

[3] H. Berestycki, L. Caffarelli, and L. Nirenberg, *Uniform estimates for regularisation of free boundary problems*, in C. Sadosky & M. Decker, eds., Anal. and Part. Diff. Eq., 1990, pp. 567–617.

[4] H. Berestycki and B. Larrouturou, *Quelques aspects mathématiques de la propagation des flammes prémélangées*, in Nonlinear P.D.E. and their Applications, Collège de France seminar, 10, Brézis & Lions, eds., Pitman Longman, Harbow, UK, 1990.

[5] H. Berestycki and B. Larrouturou, *A semilinear elliptic equation in a strip arising in a two-dimensional flame propagation model*, J. Reine Angew. Math., 396 (1989), pp. 14–40.

[6] H. Berestycki, B. Larrouturou, and P. L. Lions, *Multidimensional traveling-wave solutions of a flame propagation model*, Arch. Rational Mech. Anal., 111 (1990), pp. 33–49.

[7] H. Berestycki and L. Nirenberg, *On the method of moving planes and the sliding method*, Bol. Soc. Brasil. Mat., 22 (1991), pp. 1–37.

[8] H. Berestycki and L. Nirenberg, *Asymptotic behaviour via Harnack inequality*, in Nonlinear Analysis, A Tribute in Honour of Giovanni Prodi., Scuola Normale Superiore, Pisa, Quaderni, Univ. di Pisa, Pisa, 1991, pp. 135–144.

[9] H. Berestycki and L. Nirenberg, *Travelling fronts in cylinders*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 9 (1992), pp. 497–572.

[10] C. Bernardi and Y. Maday, *Properties of some weighted Sobolev spaces and application to spectral approximates*, SIAM J. Numer. Anal., 26 (1988), pp. 769–829.

[11] J. D. Buckmaster and G. S. S. Ludford, *Lectures on Mathematical Combustion*, CBMS-NSF Conf. Ser. Appl. Math. 43, SIAM, Philadelphia, PA, 1983.

[12] L. Caffarelli, *A Harnack inequality approach to the regularity of free boundaries, Part* II*: Flat free boundaries are Lipschitz*, Comm. Pure Appl. Math., XLII (1989), pp. 55–78.

[13] M. Dauge, *Elliptic boundary value problems on corners domains*, Lecture Notes in Math., 1341 Springer, Paris, 1988.

[14] P. C. Fife, *Mathematical aspects of reacting and diffusing systems*, Lecture Notes in Biomath. 28, Springer, New York, 1979.

[15] P. C. Fife and J. B. McLeod, *The approach of solutions of non-linear diffusion equations to traveling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[16] G. Geymonat and P. Grisvard, *Eigenfunctions expansions for non self-adjoint operators and separations of variables*, in Singularities and Constructive Methods fot their Treatment, P. Grisvard, W.L. Wendland, J.R. Whiteman, eds., Lecture Notes in Math., 1121, Springer, New York, 1985.

[17] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin, 1983.

[18] L. Glangetas and J. M. Roquejoffre, *Bifurcations of travelling waves in the thermo-diffusive model for flame propagation*, Arch. Rational Mech. Anal., 134 (1996), pp. 341–402.

[19] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, MA, 1985.

[20] P. Grisvard, *Singularities in boundary value problems*, Res. Notes Appl. Math., Springer-Verlag, Berlin, 1992.

[21] F. Hamel and R. Monneau, *Solutions of semilinear elliptic equations in $\mathbb{R}^N$ with conical-shaped level sets*, Preprint Labo. Ana. Num. Paris VI, R98029 (1998), submitted.

[22] F. Hamel and R. Monneau, *Existence and uniqueness of solutions of a conical shaped free boundary problem in $\mathbb{R}^2$*, manuscript, 1999.

[23] G. Joulin, *Dynamique des fronts de flammes*, in Modélisation de la combustion, Images des Mathématiques, CNRS, 1996 (in French).

[24] Ya. I. Kanel', *Certain problems of burning-theory equations*, Sov. Math. Dokl., 2 (1961), pp. 48–51.

[25] V. A. Kondrat'ev, *Boundary problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[26] A. N. Kolmogorov, I. G. Petrovsky, and N. S. Piskunov, *Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Bulletin Université d'Etat à Moscou (Bjul. Moskowskogo Gos. Univ.), Série internationale, section A 1 (1937), pp. 1–26 (in French).

[27] A. Liñan, *The structure of diffusion flames*, in Fluid Dynamical Aspects of Combustion Theory, Pitman Res. Notes Math. Ser. 223, Longman Sci. Tech., Harlow, 1991, pp. 11–29. S.B. MARGOLIS, Bifurcation phenomena in burner-stabilized premixed flames, Comb.

[28] S. B. Margolis and G. I. Sivashinsky, *Flame propagation in vertical channels: Bifurcation to bimodal cellular flames*, SIAM J. Appl. Math., 44 (1984), pp. 344–368.

[29] B. J. Matkowsky and G. I. Sivashinsky, *An asymptotic derivation of two models in flame theory associated with the constant density approximation*, SIAM J. Appl. Math., 37 (1979), pp. 686–699.

[30] V. G. Maz'ja and B. A. Plamenevskii, *On the coefficients in the asymptotics of solutions of elliptic boundary-value problems near conical points*, Soviet Math. Dokl., 15 (1974), pp. 1574–1575.

[31] G. I. Sivashinsky, *The structure of Bunsen flames*, J. Chem. Phys., 62 (1975), pp. 638–643.

[32] G. I. Sivashinsky, *The diffusion stratification effect in Bunsen flames*, J. Heat Transfer, Transactions of ASME, 11 (1974), pp. 530–535.

[33] J. M. Vega, *Multidimensional travelling fronts in a model from combustion theory and related problems*, Differential Integral Equations, 6 (1993), pp. 131–155.

[34] A. I. Volpert and V. A. Volpert, *Application of the Leray-Schauder degree to investigation of traveling wave solutions of parabolic systems*, in Elliptic and Parabolic Problems, Pont-à-Mousson 1994, Pitman Res. Notes in Math. Series 325, 1994, pp. 224–239.

[35] F. Williams, *Combustion Theory*, Addison-Wesley, Reading, MA, 1983.

[36] X. Xin, *Existence and uniqueness of travelling waves in a reaction-diffusion equation with combustion nonlinearity*, Idiana Univ. Math. J., 40 (1991), pp. 985–1008.

[37] J. B. Zeldovich and D. A. Frank-Kamenetskii, *A theory of thermal propagation of flame*, Acta Phys. URSS, 9 (1938), pp. 341–350 (in Russian). Dynamics of curved fronts, R. Pelcé Ed., Perspectives in Physics Series, Academic Press, New York, 1988, pp. 131–140 (in English).

SIAM J. Math. Anal.
Vol. 31, No. 1, pp. 119–133

EXISTENCE AND NONEXISTENCE OF SOLUTIONS OF NONLINEAR NEUMANN PROBLEMS[*]

STANISLAV I. POHOZAEV[†] AND ALBERTO TESEI[‡]

Abstract. Existence theorems for nonnegative solutions to a class of nonlinear Neumann problems are proved. Nonexistence results are also discussed, depending either on absorption or on first-order terms. The proofs make use of a direct variational approach.

Key words. Neumann problems, problems of indefinite type, variational methods, fibering method, nonnegative solutions, existence of solutions

AMS subject classifications. 35J20, 35J25, 35J65

PII. S0036141098334948

1. Introduction.

In this paper we study the existence of nonnegative solutions to the nonlinear Neumann problem

$$\text{(1.1)}\qquad \begin{cases} -\Delta_p u = (\nabla\psi(x), \nabla u) \mid \nabla u \mid^{p-2} + a(x)u^{q-1} + b(x)u^{s-1} & \text{in } \Omega, \\ \\ \mid \nabla u \mid^{p-2} \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

Here $\Omega \subseteq \mathbb{R}^n$ is a connected bounded domain with $C^{1,\alpha}$ boundary; by ν we denote the outer normal at any point $x \in \partial\Omega$. We also set

$$\Delta_p u \equiv \operatorname{div}(\mid \nabla u \mid^{p-2} \nabla u)$$

for $p > 1$ and

$$(\nabla\psi, \nabla u) \equiv \sum_{i=1}^{n} \frac{\partial \psi}{\partial x_i} \frac{\partial u}{\partial x_i}.$$

The functions a, b are continuous in $\bar{\Omega}$, while (ψ is differentiable) $\nabla\psi$ is bounded and uniformly continuous in Ω. An essential feature of the problem is that the function a changes sign (namely, the problem is of indefinite type; see [9]); instead, the function b is assumed to be nonpositive. Concerning the exponents q, s we shall always make the following hypothesis:

$$(\mathrm{H}_0)\qquad 1 < q < p^*, \qquad 1 < s < p^*,$$

where

$$p^* := \begin{cases} \frac{np}{n-p} & \text{if } p < n, \\ \\ \infty & \text{otherwise.} \end{cases}$$

[*] Received by the editors March 6, 1998; accepted for publication (in revised form) March 16, 1999; published electronically November 23, 1999. This work was partially supported through TMR Programme NPE FMRX-CT98-0201.

http://www.siam.org/journals/sima/31-1/33494.html

[†] Steklov Mathematical Institute, Russian Academy of Sciences, Gubkina 8, 117966 Moscow, Russia (pohozaev@mi.ras.ru).

[‡] Dipartimento di Matematica "G. Castelnuovo," Università di Roma "La Sapienza," P.le A. Moro 5, I-00185 Rome, Italy (tesei@mat.uniroma1.it).

Problem (1.1) is suggested by some mathematical models of the applied sciences (e.g., see [1], [12]); besides, in several respects it generalizes other problems previously dealt with in the literature. In particular, if $p = 2$, $\psi =$ constant, and $b \equiv 0$, it reads

$$\begin{cases} -\Delta u = a(x)\, u^{q-1} & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega. \end{cases}$$

This was investigated in [2] in the case $1 < q < 2$, respectively, and in [3] in the case $2 < q < 2^*$. In both cases the following conditions,

$$(\mathrm{H}_1) \qquad a_+ := \max\{a, 0\} \not\equiv 0,$$

$$(1.2) \qquad \int_\Omega a\, dx < 0$$

are necessary and sufficient for the existence of positive solutions.

As already mentioned, we retain (H$_1$) in the present investigation. As for the latter condition, recasting (1.1) in the equivalent form

$$(1.3) \qquad \begin{cases} -\mathrm{div}\,(\rho(x) \mid \nabla u \mid^{p-2} \nabla u) = \rho(x)a(x)u^{q-1} + \rho(x)b(x)u^{s-1} & \text{in } \Omega, \\ \mid \nabla u \mid^{p-2} \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \end{cases}$$

where $\rho := e^\psi$, suggests the more general assumption

$$(\mathrm{H}_2) \qquad \int_\Omega \rho\, a\, dx < 0.$$

In the following we always assume (H$_0$)–(H$_2$) and moreover

$$(\mathrm{H}_3) \qquad b \leq 0 \quad \text{in } \Omega.$$

When $b \equiv 0$ the following theorem holds, which generalizes previous existence results in [2], [3].

Theorem 1.1. *Let assumptions* (H$_0$)–(H$_2$) *be satisfied; let $b \equiv 0$. Then there exists a nontrivial nonnegative solution $u \in L^\infty(\Omega)$ of problem* (1.1). *Moreover, $u \in C^{1,\beta}(\bar{\Omega})$ for some $\beta > 0$.*

Under the more general assumption (H$_3$) the relationship between the exponents p, q, and s plays an essential role. If either

$$(\mathrm{A}) \qquad q > \max\{p, s\}$$

or

$$(\mathrm{B}) \qquad q < \min\{p, s\},$$

the following result applies.

Theorem 1.2. *Let either* (A) *or* (B) *hold and assumptions* (H$_0$)–(H$_3$) *be satisfied. Then there exists a nontrivial nonnegative solution $u \in L^\infty(\Omega)$ of problem* (1.1). *Moreover, $u \in C^{1,\beta}(\bar{\Omega})$ for some $\beta > 0$.*

The remaining cases, namely,

$$s < q < p, \tag{C}$$

$$p < q < s, \tag{D}$$

are more cumbersome. Assuming that

$$\operatorname{supp} a_+ \setminus \operatorname{supp} b \quad \text{has nonempty interior,} \tag{H$_4$}$$

for case (C), the following can be proved.

THEOREM 1.3. *Let* (C) *hold and assumptions* (H_0)–(H_4) *be satisfied. Then there exists a nontrivial nonnegative solution $u \in L^\infty(\Omega)$ of problem* (1.1). *Moreover, $u \in C^{1,\beta}(\bar{\Omega})$ for some $\beta > 0$.*

To deal with case (D) we shall use the following assumption:

$$b(x) \leq -b_0 < 0 \quad \text{for any x} \in \Omega. \tag{H$_5$}$$

(Observe that conditions (H_4) and (H_5) exclude each other.) In the following statement, by saying that "*a_+ is large with respect to b*," we mean that condition (H_6) below (see section 2) is satisfied.

THEOREM 1.4. *Let* (D) *hold and assumptions* (H_0)–(H_2) *and* (H_5) *be satisfied. Assume that a_+ is large with respect to b. Then there exists a nontrivial nonnegative solution $u \in L^\infty(\Omega)$ of problem* (1.1). *Moreover, $u \in C^{1,\beta}(\bar{\Omega})$ for some $\beta > 0$.*

According to the above theorem, in case (D) a nontrivial, nonnegative solution exists if the source term $a_+(x)u^{q-1}$ prevails over the absorption term $b(x)u^{s-1}$. In the opposite case such a solution does not exist, as the following result shows. By saying that "*b is large with respect to a_+*," we mean that condition (H_7) (see section 4) is satisfied.

THEOREM 1.5. *Let* (D) *hold and assumptions* (H_0)–(H_2) *and* (H_5) *be satisfied. Assume that b is large with respect to a_+. Then the only nonnegative solution of problem* (1.1) *is trivial.*

Observe that the above nonexistence result depends (for fixed functions a, ρ) on the magnitude of the absorption coefficient b. A different nonexistence result, which depends only on the first-order term, can be pointed out. If $b \equiv 0$ and $q > 2$, condition (H_2) is necessary for the existence of a nontrivial, nonnegative solution to (1.1) (see Proposition 4.1). Suppose that condition (1.2) is satisfied, while (H_2) is not. In this case nontrivial, nonnegative solutions of (1.1) (with $b \equiv 0$, $q > 2$) exist if $\psi =$ constant by Theorem 1.1, yet they do not exist for general ψ. Similar nonexistence phenomena due to the effect of first-order terms are known for Dirichlet boundary value problems and for free boundary problems (see [4], [5]).

The proofs of the above results make use of direct variational arguments introduced in [13], [14], [15] (see also [8]); an outline is given in section 2 for convenience of the reader.

2. Mathematical framework and results. Let X be a real Banach space with norm $\|\cdot\|$; let f and H be real-valued functionals defined in X. Let f, H be continuously differentiable in $X \setminus \{0\}$; suppose that $H(0) = 0$ and

$$\langle H'(v), v\rangle \neq 0 \tag{2.1}$$

for any v such that

$$H(v) = 1. \tag{2.2}$$

Here H' denotes the derivative of H and $\langle \cdot, \cdot \rangle$ denotes the pairing between X and its dual space.

We associate with f a functional F setting

$$F(r, v) := f(rv) \tag{2.3}$$

for any $r \in \mathbb{R}$ and $v \in X$.

Proposition 2.1. *Let (r, v) be a conditionally critical point of F under condition (2.2) such that $r \neq 0$. Then $u := rv$ is a nonzero critical point of the functional f.*

Proof. According to the rule of Lagrange multipliers, there exist λ, $\mu \in \mathbb{R}$ such that $\lambda^2 + \mu^2 > 0$ and

$$\lambda F_v(r, v) = \mu H'(v), \tag{2.4}$$

$$\lambda F_r(r, v) = 0. \tag{2.5}$$

Here F_r, F_v denote the partial derivatives of F. By the definition (2.3) of F we have

$$rF_r(r, v) = \langle F_v(r, v), v \rangle \tag{2.6}$$

and

$$f'(u) = \frac{1}{r} F_v \left(r, \frac{u}{r} \right). \tag{2.7}$$

Since $\langle H'(v), v \rangle \neq 0$ by assumption, we obtain $\lambda \neq 0$, $\mu = 0$. Then the conclusion follows. □

Suppose that in some open subset $E \subseteq X \setminus \{0\}$ a real-valued, continuously differentiable functional $r = r(v)$ is defined, such that $r(v) \neq 0$ and

$$F_r(r(v), v) = 0 \tag{2.8}$$

for any $v \in E$ such that condition (2.2) is satisfied. Define a functional $\tilde{f}(v)$ setting

$$\tilde{f}(v) := F(r(v), v). \tag{2.9}$$

Proposition 2.2. *Let v be a conditionally extremum point of $\tilde{f}(v)$ under condition (2.2). Then $u := r(v)v$ is a nonzero critical point of the functional f.*

Proof. If $r = r(v)$, equality (2.8) and the definition of $\tilde{f}$ ensure that (r, v) is a conditionally critical point of F under condition (2.2). Then the conclusion follows by Proposition 2.1. □

The previous results suggest the following approach to investigating critical points of the functional f. First we study the equation

$$F_r(r, v) = 0, \tag{2.10}$$

referred to as the *bifurcation equation*. Suppose that for any v in some open subset $E \subseteq X \setminus \{0\}$ there exists a root $r = r(v) \neq 0$ of (2.10); let $r \in C^1(E)$. Then the *reduced functional* $\tilde{f}$ given by (2.9) is defined and is of class C^1 in E; following Proposition

2.2, we maximize (or minimize) $\tilde{f}$ under the constraint $H(v) = 1$, where H is some suitable functional.

Let us investigate problem (1.3) using the previous considerations. We shall work in the Sobolev space $X = W^{1,p}(\Omega)(1 < p < \infty)$ endowed with the norm

$$|u|_X := \left\{ \int_\Omega \rho\, |u|^p \, dx + \int_\Omega \rho\, |\nabla u|^p \, dx \right\}^{1/p} .$$

Since ρ is bounded away from zero in Ω, this norm is equivalent to the usual one. The functional f associated with (1.3) is

$$f(u) = -\frac{1}{p} \int_\Omega \rho\, |\nabla u|^p \, dx + \frac{1}{q} \int_\Omega \rho\, a\, |u|^q \, dx + \frac{1}{s} \int_\Omega \rho\, b\, |u|^s \, dx.$$

The functional (2.3) and the bifurcation equation (2.10) read in the present case

$$F(r, v) = -\frac{|r|^p}{p} \int_\Omega \rho\, |\nabla v|^p \, dx + \frac{|r|^q}{q} A(v) - \frac{|r|^s}{s} B(v),$$

respectively,

$$F_r(r, v) = A(v)|r|^{q-2}r - B(v)|r|^{s-2}r - \int_\Omega \rho\, |\nabla v|^p \, dx\, |r|^{p-2}r = 0; \tag{2.11}$$

here

$$A(v) := \int_\Omega \rho\, a\, |v|^q \, dx,$$

$$B(v) := \int_\Omega \rho\, |b|\, |v|^s \, dx,$$

and use of the assumption (H_3) has been made.

For $r \neq 0$ the bifurcation equation (2.11) is equivalent to

$$\phi(r, v) = \int_\Omega \rho\, |\nabla v|^p \, dx,$$

where

$$\phi(r, v) := A(v)|r|^{q-p} - B(v)|r|^{s-p}. \tag{2.12}$$

Set

$$E := \{v \in X \mid A(v) > 0\}; \tag{2.13}$$

observe that by assumption (H_1) the set E is nonempty.

It is apparent from (2.12) that, if $b \equiv 0$ in Ω, for any $v \in E$ the bifurcation equation has a unique positive root, namely,

$$r(v) = \left\{ \frac{\int_\Omega \rho\, |\nabla v|^p \, dx}{A(v)} \right\}^{1/(q-p)} .$$

Then the reduced functional reads

$$\tilde{f}(v) = \left(\frac{1}{q} - \frac{1}{p}\right) \left\{ \frac{[\int_\Omega \rho\,|\nabla v|^p\,dx]^q}{[A(v)]^p} \right\}^{1/(q-p)} . \tag{2.14}$$

In the following we always choose the functional H as follows:

$$H(v) := \int_\Omega \rho\,|\nabla v|^p\,dx.$$

Concerning the variational problem

$$\max_{v \in E} \tilde{f}(v) \quad \text{under the condition} \quad \int_\Omega \rho\,|\nabla v|^p\,dx = 1, \tag{2.15}$$

the following result will be proved.

Proposition 2.3. *Let assumptions* (H$_0$)–(H$_2$) *be satisfied; let $b \equiv 0$. Then the maximum in* (2.15) *(where $\tilde{f}$ is the functional* (2.14)*) is achieved at some function $\bar{v} \geq 0$, $\bar{v} \not\equiv 0$ in Ω.*

Theorem 1.1 is an immediate consequence of the above proposition. Similarly, Theorems 1.2–1.4 follow easily from Propositions 2.4–2.6 below.

If $b \not\equiv 0$ and either (A) or (B) holds, the bifurcation equation has a unique positive root $r = r(v)$ for any $v \in E$ (see (2.12)). Moreover, for any $v \in E$ the quantity

$$\langle f''(r(v)v)v, v\rangle = F_{rr}(r(v), v)$$

$$= (q-p) \int_\Omega \rho\,|\nabla v|^p\,dx\,|r(v)|^{p-2} + (q-s)B(v)\,|r(v)|^{s-2}$$

is strictly positive if (A) holds or negative if (B) holds; hence $r \in C^1(E)$. The reduced functional

$$\tilde{f}(v) = \left(\frac{1}{q} - \frac{1}{p}\right) \int_\Omega \rho\,|\nabla v|^p\,dx\,|r(v)|^p + \left(\frac{1}{q} - \frac{1}{s}\right) B(v)\,|r(v)|^s$$

$$= \left(\frac{1}{q} - \frac{1}{p}\right) A(v)\,|r(v)|^q - \left(\frac{1}{s} - \frac{1}{p}\right) B(v)\,|r(v)|^s \tag{2.16}$$

is defined for any $v \in E$; the following result will be proved.

Proposition 2.4. *Let assumptions* (H$_0$)–(H$_3$) *be satisfied; moreover, let either* (A) *or* (B) *hold. Then the maximum in* (2.15) *(where $\tilde{f}$ is the functional* (2.16)*) is achieved at some function $\bar{v} \geq 0$, $\bar{v} \not\equiv 0$ in Ω.*

Concerning case (C), it is easily seen from (2.12) that a nontrivial solution $r(v)$ of the bifurcation equation exists for any $v \in E$ such that $B(v) = 0$, yet need not exist if $B(v) > 0$. In the latter case the function $\phi(\cdot, v)$ has a unique positive maximum point, namely,

$$r_*(v) := \left\{ \frac{p-s}{p-q}\,\frac{B(v)}{A(v)} \right\}^{1/(q-s)} . \tag{2.17}$$

Moreover,

$$\phi(r_*(v), v) = \left\{ \frac{[A(v)]^{p-s}}{\gamma_0 [B(v)]^{p-q}} \right\}^{1/(q-s)} ,$$

where

$$\gamma_0 := \frac{(p-s)^{p-s}}{(q-s)^{q-s}(p-q)^{p-q}}. \tag{2.18}$$

Define

$$E_0 := \left\{ v \in E \mid [A(v)]^{p-s} > \gamma_0 \, [B(v)]^{p-q} \left[\int_\Omega \rho \, |\nabla v|^p \, dx \right]^{q-s} \right\}.$$

Observe that $E_0 \neq \emptyset$ by assumption (H_4); in fact, for any $v \neq 0$ such that

$$\operatorname{supp} v \subseteq (\operatorname{supp} a_+ \setminus \operatorname{supp} b)^o$$

we have $A(v) > 0$, $B(v) = 0$.

For any $v \in E_0$ the bifurcation equation has one positive root if $B(v) = 0$ or two, say,

$$r_-(v) < r_*(v) < r_+(v),$$

if $B(v) > 0$. In both cases we denote by $r(v)$ the maximal positive root. Observe that the quantity $F_{rr}(r(v), v)$ is strictly negative; hence $r \in C^1(E_0)$; in fact,

$$F_{rr}(r(v), v) = -(p-q)A(v) \, |r(v)|^{q-2} \quad \text{if } B(v) = 0$$

or

$$F_{rr}(r(v), v) = -(p-q)A(v) \, |r(v)|^{s-2}[|r(v)|^{q-s} - |r_*(v)|^{q-s}] \quad \text{if } B(v) > 0.$$

Since the reduced functional (2.16) is defined for any $v \in E_0$, the variational problem

$$\max_{v \in E_0} \tilde{f}(v) \quad \text{under the condition} \quad \int_\Omega \rho \, |\nabla v|^p \, dx = 1 \tag{2.19}$$

can be investigated. The following proposition holds.

Proposition 2.5. *Let assumptions* (H_0)–(H_4) *be satisfied; let* (C) *hold. Then the maximum in* (2.19) *is achieved at some function* $\bar{v} \geq 0$, $\bar{v} \not\equiv 0$ *in* Ω.

Finally, let us discuss case (D) under assumption (H_5); observe that this assumption implies $B(v) > 0$ whenever $A(v) > 0$. Instead of the set E_0 considered in case (C), now define

$$E_1 := \left\{ v \in E \mid [A(v)]^{s-p} > \gamma_1 \, [B(v)]^{q-p} \left[\int_\Omega \rho \, |\nabla v|^p \, dx \right]^{s-q} \right\},$$

where

$$\gamma_1 := \frac{(s-p)^{s-p}}{(s-q)^{s-q} \, (q-p)^{q-p}}. \tag{2.20}$$

If $E_1 \neq \emptyset$, for any $v \in E_1$ there exist two positive solutions of the bifurcation equation. In such a case we denote again by $r(v)$ the maximal positive root and consider the reduced functional (2.16) for $v \in E_1$.

To ensure that the set E_1 be nonempty, we find it convenient to introduce its subset

$$E_2 := \left\{ v \in E \mid [A(v)]^{s-p} > \gamma_2 \, [B(v)]^{q-p} \left[\int_\Omega \rho \, |\nabla v|^p \, dx \right]^{s-q} \right\},$$

where

$$\gamma_2 := \frac{q^{s-p}}{p^{s-q} \, s^{q-p}} \, \gamma_1.$$

It is easily checked that $\gamma_2 > \gamma_1$; thus $E_2 \subseteq E_1$ as asserted. We shall assume that

$$(\mathrm{H}_6) \qquad E_2 \text{ is nonempty.}$$

It is easily proven that for any $v \in E_2$ the functional $F(\cdot, v)$ has two positive zeros. Hence it has a (local) minimum point and a maximum point, which are the minimal, respectively, the maximal positive root of the bifurcation equation. It follows that

$$\tilde{f}(v) = F(r(v), v) = \max_{r>0} F(r, v) > 0$$

for any $v \in E_2$.

Concerning the problem

$$(2.21) \qquad \max_{v \in E_1} \; \tilde{f}(v) \quad \text{under the condition} \quad \int_\Omega \rho \, |\nabla v|^p \, dx = 1,$$

the following result will be proved.

PROPOSITION 2.6. *Let assumptions* (H$_0$)–(H$_2$) *and* (H$_5$)–(H$_6$) *be satisfied; let* (D) *hold. Then the maximum in* (2.21) *is achieved at some function* $\bar{v} \geq 0$, $\bar{v} \not\equiv 0$ *in* Ω.

In connection with assumption (H$_6$) and the statement of Theorem 1.4, observe that the inequality

$$[A(v)]^{s-p} > \gamma_2 \, [B(v)]^{q-p} \left[\int_\Omega \rho \, |\nabla v|^p \, dx \right]^{s-q}$$

is satisfied for some $v \in E$ if the positive part a_+ is sufficiently large with respect to b (for instance, it suffices to replace a_+ by λa_+, $\lambda > 0$ large enough).

3. Proofs of existence. Set

$$S := \left\{ v \in X | \int_\Omega \rho \, |\nabla v|^p \, dx = 1 \right\}.$$

Let us prove the following lemma.

LEMMA 3.1. *Let assumptions* (H$_0$)–(H$_2$) *be satisfied. Then the set*

$$E \cap S = \left\{ v \in X | A(v) > 0, \int_\Omega \rho \, |\nabla v|^p \, dx = 1 \right\}$$

is bounded in X.

Proof. By absurd, let $\{v_n\} \subseteq E \cap S$ be such that

$$\int_\Omega \rho \, |v_n|^p \, dx + \int_\Omega \rho \, |\nabla v_n|^p \, dx \longrightarrow \infty$$

as $n \to \infty$. For any $n \in \mathbb{N}$ set

$$v_n = t_n + w_n,$$

where

$$t_n := \frac{1}{\|\rho\|_1} \int_\Omega \rho v_n dx,$$

$$w_n := v_n - t_n.$$

Since

$$\int_\Omega \rho \, |\nabla w_n|^p \, dx = \int_\Omega \rho \, |\nabla v_n|^p \, dx = 1$$

and

$$\int_\Omega \rho \, w_n \, dx = 0,$$

by embedding results there exists $C > 0$ such that

$$|w_n|_X \leq C \quad \text{for any } n \in \mathbb{N}.$$

This implies that $|t_n| \to \infty$; moreover, since by assumption (H_0) the space X is compactly embedded in $L^q(\Omega)$, we may assume that $\{w_n\}$ converges strongly in the latter space. Then we have

$$\int_\Omega \rho \, a \, |v_n|^q \, dx = |t_n|^q \int_\Omega \rho \, a \left| 1 + \frac{w_n}{t_n} \right|^q \, dx \longrightarrow -\infty$$

as $n \to \infty$ by assumption (H_2). This contradicts the definition of E; hence the conclusion follows. □

Let us prove Proposition 2.4. The proof of Proposition 2.3 is similar, yet simpler by the homogeneity of the reduced functional (2.14); hence it is omitted.

Proof of Proposition 2.4. Set

$$M := \sup\{\tilde{f}(v) \mid v \in E \cap S\}, \tag{3.1}$$

where $\tilde{f}$ is the reduced functional (2.16). It is easily seen that $M \in (-\infty, 0]$ if (A) holds or $M \in (0, \infty)$ if (B) is satisfied. Let $\{v_n\} \subseteq E \cap S$ be a maximizing sequence. Due to Lemma 3.1, we can assume that $\{v_n\}$ converges weakly in X to some $\bar{v}$; by assumption (H_0), it follows that $v_n \to \bar{v}$ both in $L^q(\Omega)$ and in $L^s(\Omega)$. Let us prove that $\bar{v} \in E \cap S$.

(i) Since $\{v_n\} \subseteq E \cap S$, from the bifurcation equation we obtain

$$A(v_n)|r(v_n)|^{q-p} \geq 1 \quad \text{for any } n \in \mathbb{N}. \tag{3.2}$$

On the other hand, since $v_n \to \bar{v}$ in $L^q(\Omega)$, there holds

$$A(v_n) \to A(\bar{v}) \quad \text{as } n \to \infty.$$

By absurd, let $A(\bar{v}) = 0$. If (A) holds, let us rewrite (3.2) as follows:

$$|r(v_n)| \geq [A(v_n)]^{-1/(q-p)};$$

then we conclude that $|r(v_n)| \to \infty$. Since by (2.16)

$$\tilde{f}(v_n) \le \left(\frac{1}{q} - \frac{1}{p}\right) |r(v_n)|^p,$$

this implies that $\tilde{f}(v_n) \to -\infty$, which is impossible. If (B) holds, we can recast (3.2) in the following form,

$$A(v_n) \ge |r(v_n)|^{p-q}, \tag{3.3}$$

thus obtaining that $|r(v_n)| \to 0$. Since $B(v_n) \to B(\bar{v}) < \infty$, this implies that $\tilde{f}(v_n) \to 0$, contradicting the inequality $M > 0$. Then $A(\bar{v}) > 0$, i.e., $\bar{v} \in E$.

(ii) By the weak convergence of $\{v_n\}$ in X there holds

$$\int_\Omega \rho\, |\nabla \bar{v}|^p \, dx \le 1.$$

Since $A(\bar{v}) > 0$, by (H_2) we also have

$$\int_\Omega \rho\, |\nabla \bar{v}|^p \, dx > 0.$$

If the first inequality were strict, we could find $t > 1$ such that

$$\int_\Omega \rho |\nabla (t\bar{v})|^p dx = 1;$$

hence $t\bar{v} \in E \cap S$. The root $r = r(t\bar{v})$ of the bifurcation equation satisfies the equality

$$A(t\bar{v})\, |r(t\bar{v})|^{q-p} - B(t\bar{v})\, |r(t\bar{v})|^{s-p} = 1. \tag{3.4}$$

Since

$$A(t\bar{v}) = t^q\, A(\bar{v}),$$

$$B(t\bar{v}) = t^s\, B(\bar{v}),$$

this gives

$$A(\bar{v})\, |tr(t\bar{v})|^{q-p} - B(\bar{v})\, |tr(t\bar{v})|^{s-p} = t^{-p} < 1. \tag{3.5}$$

On the other hand, it is easily seen that the sequence $\{r(v_n)\}$ is bounded. In fact, in case (B) this follows from inequality (3.3). Concerning (A), rewrite the bifurcation equation for $v = v_n$ as

$$|r(v_n)|^{q-p} \{A(v_n) - B(v_n) |r(v_n)|^{s-q}\} = 1.$$

Since $A(v_n) \to A(\bar{v}) > 0$ and $\{B(v_n)\}$ is converging, for any diverging subsequence of $\{r(v_n)\}$ the left-hand side of the above equality would diverge, which is impossible. Since $\{r(v_n)\}$ is bounded, some subsequence is converging; then its limit, say, $\bar{r}$, satisfies the equality

$$A(\bar{v})\, |\bar{r}|^{q-p} - B(\bar{v})\, |\bar{r}|^{s-p} = 1. \tag{3.6}$$

Comparing (3.5) and (3.6) immediately gives

$$tr(t\bar{v}) < \bar{r}$$

if (A) holds, respectively,

$$tr(t\bar{v}) > \bar{r}$$

if (B) is satisfied. Then an elementary investigation of the function

$$\psi(\xi) := \left(\frac{1}{q} - \frac{1}{p}\right) A(\bar{v})\, \xi^q - \left(\frac{1}{s} - \frac{1}{p}\right) B(\bar{v})\, \xi^s \qquad (\xi > 0) \tag{3.7}$$

proves that in both cases

$$\tilde{f}(t\bar{v}) = \psi(t|r(t\bar{v})|) > \psi(\bar{r}) = M, \tag{3.8}$$

which is absurd. It follows that $\bar{v} \in S$; thus the claim is proved. Since the equality (3.4) holds with $t = 1$, we get $r(\bar{v}) = \bar{r}$ (see (3.6)); thus $M = \tilde{f}(\bar{v})$. Then the conclusion follows. □

Let us now consider case (C).

Proof of Proposition 2.5. Set

$$M := \sup\{\tilde{f}(v) \mid v \in E_0 \cap S\}.$$

Observe that $M > 0$ by assumption (H_4); in fact, for any $v \not\equiv 0$ with

$$\operatorname{supp} v \subseteq (\operatorname{supp} a_+ \setminus \operatorname{supp} b)^o$$

there holds

$$\tilde{f}(v) = \left(\frac{1}{q} - \frac{1}{p}\right) A(v)\, |r(v)|^q > 0.$$

Since $E_0 \cap S \subseteq E \cap S$ is bounded in X (see Lemma 3.1), any maximizing sequence $\{v_n\} \subseteq E_0 \cap S$ converges to some $\bar{v} \in X$ as in the proof of Proposition 2.4. Let us prove that $\bar{v} \in E_0 \cap S$.

To this purpose, observe first that the sequence $\{r(v_n)\}$ is bounded. In fact, for any diverging subsequence the right-hand side of the equality

$$A(v_n)\, |r(v_n)|^{q-p} - B(v_n)\, |r(v_n)|^{s-p} = 1 \tag{3.9}$$

would be infinitesimal, which is impossible.

(i) Let us show that $A(\bar{v}) > 0$. Since $\{r(v_n)\}$ is bounded, some subsequence (again denoted $\{r(v_n)\}$) converges to a limit $\bar{r}$. If $A(\bar{v}) = 0$, we have

$$M = \lim_{n\to\infty} \tilde{f}(v_n) = \left(\frac{1}{p} - \frac{1}{s}\right) B(\bar{v}) |\bar{r}|^s \leq 0,$$

which is absurd since $M > 0$.

(ii) Let us prove that the strict inequality in the definition of the set E_0 is satisfied at $v = \bar{v}$. This follows by (i) if $B(\bar{v}) = 0$. In any case there holds

$$[A(\bar{v})]^{p-s} \geq \gamma_0\, [B(\bar{v})]^{p-q} \left[\int_\Omega \rho\, |\nabla \bar{v}|^p\, dx\right]^{q-s}.$$

Suppose that $B(\bar{v}) > 0$ and the equality sign holds in the above relation. This means that the maximum of the function $\phi(\cdot, \bar{v})$ equals $\int_\Omega \rho\, |\nabla \bar{v}|^p\, dx$ (see section 2); hence

$$r(\bar{v}) = r_*(\bar{v}).$$

On the other hand, it follows easily from (3.9) that $\bar{r} := \lim_{n\to\infty} |r(v_n)|$ is strictly positive. Passing to the limit in the same equation as $n \to \infty$ we find

$$A(\bar{v})\, |\bar{r}|^{q-p} - B(\bar{v})\, |\bar{r}|^{s-p} = 1 \geq \int_\Omega \rho\, |\nabla \bar{v}|^p\, dx, \tag{3.10}$$

whence

$$\bar{r} = r(\bar{v}) = r_*(\bar{v}).$$

Then we obtain

$$M = \lim_{n\to\infty} \tilde{f}(v_n)$$

$$= \left(\frac{1}{q} - \frac{1}{p}\right) A(\bar{v})\, r^*(\bar{v})^q + \left(\frac{1}{p} - \frac{1}{s}\right) B(v)\, r^*(\bar{v})^s$$

$$= \frac{1}{p}\left(\frac{1}{q} - \frac{1}{s}\right)\left\{\frac{[(p-s)B(\bar{v})]^q}{[(p-q)A(\bar{v})]^s}\right\}^{1/(q-s)} < 0,$$

which is absurd. Hence the claim follows.

(iii) It is easily checked that, if

$$\int_\Omega \rho\, |\nabla \bar{v}|^p\, dx < 1,$$

we can find $t > 1$ such that $t\bar{v} \in E_0 \cap S$. As in the proof of Proposition 2.4 this gives inequality (3.5). Moreover, observe that

$$r_*(\bar{v}) = t\frac{1}{t} r_*(\bar{v}) = t r_*(t\bar{v}) < t r(t\bar{v});$$

similarly,

$$r_*(\bar{v}) < \bar{r}.$$

Thus comparing (3.5) and (3.10) we find

$$\bar{r} < t r(t\bar{v}),$$

whence the claim follows as in the proof of Proposition 2.4. Since $\bar{v} \in E_0 \cap S$ and $\bar{r} = r(\bar{v})$, we have $M = \tilde{f}(\bar{v})$. Then the conclusion follows. □

Let us now prove Proposition 2.6.

Proof of Proposition 2.6. Set

$$M := \sup\{\tilde{f}(v) \mid v \in E_1 \cap S\};$$

observe that $M > 0$ by assumption (H$_6$), since $\tilde{f} > 0$ for any $v \in E_2$ (see section 2). Let $\{v_n\} \subseteq E_1 \cap S \subseteq E \cap S$ be a maximizing sequence, which converges to $\bar{v} \in X$ as in the proof of Proposition 2.4. The conclusion will follow if we prove that $\bar{v} \in E_1 \cap S$.

(i) Let us first show that $A(\bar{v}) > 0$. Since $\{v_n\} \subseteq E_1 \cap S$, for any $n \in N$ we have

$$\gamma_1 \, [B(v_n)]^{q-p} < [A(v_n)]^{s-p}$$

$$\le \left(\frac{\|a\|_\infty}{b_0} \right)^{s-p} (\|b\|_\infty \|\rho\|_1)^{(s-p)(s-q)/s} [B(v_n)]^{(s-p)q/s}. \tag{3.11}$$

Taking the limit as $n \to \infty$ gives

$$B(\bar{v}) \ge \frac{\gamma_1^{s/[p(s-q)]}}{(\|b\|_\infty \|\rho\|_1)^{(s-p)/p}} \left(\frac{b_0}{\|a\|_\infty} \right)^{[s(s-p)]/[p(s-q)]} > 0.$$

On the other hand, from the first inequality in (3.11) we obtain

$$A(\bar{v}) \ge \gamma_1^{1/(s-p)} [B(\bar{v})]^{(q-p)/(s-p)};$$

hence the claim follows.

(ii) It is easily seen that

$$[A(\bar{v})]^{s-p} \ge \gamma_1 [B(\bar{v})]^{q-p} \left[\int_\Omega \rho \, |\nabla \bar{v}|^p \, dx \right]^{s-q}.$$

To exclude the equality in the above relation we can use the same argument as in part (ii) of the proof of Proposition 2.4, provided that the sequence $\{r(v_n)\}$ is bounded. This follows easily from the bifurcation equation

$$|r(v_n)|^{s-p} \{A(v_n) |r(v_n)|^{q-s} - B(v_n)\} = 1,$$

since $B(v_n) \to B(\bar{v}) > 0$ by (i) above. Hence the claim follows.

(iii) To prove that $\int_\Omega \rho |\nabla \bar{v}|^p dx = 1$ we can use the same argument as in the proof of Proposition 2.4; we omit the details. The proof is complete. □

Proof of Theorem 1.1. Due to Proposition 2.3, there exists a conditionally extremum point $\bar{v} \ge 0$, $\bar{v} \ne 0$ of the reduced functional $\tilde{f}$ in the set E. According to Proposition 2.2 $u := r(\bar{v})\bar{v}$ is a nonzero critical point of the functional f, hence a nontrivial, nonnegative solution of problem (1.1). A standard bootstrap argument (see [7]) shows that $u \in L^\infty(\Omega)$; then the asserted regularity of u follows by [10] (see also [6], [16], [11]). Hence we have the conclusion. □

Theorems 1.2–1.4 follow similarly by Propositions 2.1 and 2.4–2.6; the details are omitted.

4. Nonexistence results. Let us briefly discuss the nonexistence results mentioned in section 1.

Proof of Theorem 1.5. Since the set $E \cap S$ is bounded in X (see Lemma 3.1), by assumptions (H$_0$) and (H$_2$) there exists $M_0 > 0$ (depending on a, ρ) such that

$$\int_\Omega \rho |v|^s dx \le M_0 \quad \text{for any } v \in E \cap S. \tag{4.1}$$

We shall prove the following statement: let

$$\|a_+\|_\infty^{s-p} \, \|\rho\|_1^{(s-p)(s-q)/s} \, M_0^{p(s-q)/s} < \gamma_1 b_0^{q-p}, \tag{H$_7$}$$

the constant γ_1 being defined in (2.20). Then the inequality

$$[A(v)]^{s-p} < \gamma_1 \, [B(v)]^{q-p} \left[\int_\Omega \rho \, |\nabla v|^p \, dx \right]^{s-q} \tag{4.2}$$

holds for any $v \in E$.

The above statement implies that the only solution of the bifurcation equation is trivial; thus the conclusion follows.

(i) Let us prove first the above statement for $v \in E \cap S$. Since

$$A(v) \leq \|a_+\|_\infty \, \|\rho\|_1^{(s-q)/s} \left(\int_\Omega \rho \, |v|^s \, dx \right)^{q/s},$$

$$B(v) \geq b_0 \int_\Omega \rho |v|^s dx,$$

the inequality (4.2) holds if

$$\|a_+\|_\infty^{s-p} \, \|\rho\|_1^{(s-p)(s-q)/s} \left(\int_\Omega \rho \, |v|^s \, dx \right)^{[p(s-q)]/s} < \gamma_1 b_0^{q-p}.$$

Due to (4.1), the latter inequality is satisfied if (H$_7$) holds. Then the conclusion will follow in this case.

(ii) By absurd, let $v \in E$ satisfy

$$[A(v)]^{s-p} \geq \gamma_1 [B(v)]^{q-p} \left[\int_\Omega \rho \, |\nabla v|^p \, dx \right]^{s-q}. \tag{4.3}$$

Recall that by (H$_2$)

$$\int_\Omega \rho \, |\nabla v|^p \, dx > 0;$$

then there exists $t > 0$ such that $tv \in E \cap S$. It is easily checked that (4.3) implies

$$[A(tv)]^{s-p} \geq \gamma_1 \, [B(tv)]^{q-p},$$

thus contradicting (i) above. This completes the proof. □

Let us finally prove the following result.

Proposition 4.1. *Let $b \equiv 0$; let $u \geq 0$, $u \neq 0$ be a solution of problem* (1.1). *Then*

$$\int_{\operatorname{supp} u} \rho \, a \, dx < 0.$$

Proof. Following [2] we set for any $\epsilon > 0$

$$h_\epsilon(s) := \int_0^s \frac{dt}{(t+\epsilon)^{(q-1)/(p-1)}}.$$

Then by (1.1) with $b \equiv 0$ the function $h_\epsilon(u(x))$ satisfies the problem

$$-\text{div} \left(\rho(x) \mid \nabla[h_\epsilon(u)] \mid^{p-2} \nabla[h_\epsilon(u)] \right)$$

$$= \rho(x)a(x)\left(\frac{u}{u+\epsilon}\right)^{q-1} + (q-1)\rho(x)\frac{|\nabla u|^p}{(u+\epsilon)^q} \quad \text{in } \Omega,$$

$$|\nabla[h_\epsilon(u)]|^{p-2}\,\frac{\partial[h_\epsilon(u)]}{\partial\nu} = \left(\frac{1}{u+\epsilon}\right)^{q-1} |\nabla u|^{p-2}\,\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega.$$

Hence

$$\int_\Omega \rho\, a \left(\frac{u}{u+\epsilon}\right)^{q-1} dx = -(q-1)\int_\Omega \rho\,\frac{|\nabla u|^p}{(u+\epsilon)^q}\,dx < 0.$$

Letting $\epsilon \to 0$ we obtain

$$\int_{\operatorname{supp} u} \rho\, a\, dx \leq 0.$$

As in [2] it is proved that the above inequality is strict; then the conclusion follows. □

Acknowledgments. Useful discussions with Prof. M. A. Pozio are gratefully acknowledged. Thanks are also due to one of the referees for drawing references [6], [10], and [11] to the authors' attention.

REFERENCES

[1] H. Amann and M. Renardy, *Reaction-diffusion problems in electrolysis*, NoDEA Nonlinear Differential Equations Appl., 1 (1994), pp. 91–117.

[2] C. Bandle, M. A. Pozio, and A. Tesei, *Existence and uniqueness of solutions of nonlinear Neumann problems*, Math. Z., 199 (1988), pp. 257–278.

[3] H. Berestycki, I. Capuzzo-Dolcetta, and L. Nirenberg, *Variational methods for indefinite superlinear homogeneous elliptic problems*, NoDEA Nonlinear Differential Equations Appl., 2 (1995), pp. 553–572.

[4] T. F. Chen, H. A. Levine, and P. E. Sacks, *Analysis of a convective reaction-diffusion equation*, Nonlinear Anal., 12 (1988), pp. 1349–1370.

[5] S. Claudi, L. A. Peletier, and A. Tesei, *A nonlinear diffusion equation involving convection and singular absorption*, J. Math. Anal. Appl., 239 (1999).

[6] E. DiBenedetto, *$C^{1+\alpha}$ Local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Anal., 7 (1983), pp. 827–850.

[7] P. Drábek, *Strongly nonlinear degenerate and singular elliptic problems*, in Nonlinear Partial Differential Equations (Fes, 1994), Pitman Res. Notes Math. Ser. 343, Longman, Harlow, UK, 1996, pp. 112–146.

[8] P. Drábek and S. I. Pohozaev, *Positive solutions for the p-Laplacian: Application of the fibering method*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 703–726.

[9] P. Hess and T. Kato, *On some linear and nonlinear eigenvalue problems with an indefinite weight function*, Comm. Partial Differential Equations, 5 (1980), pp. 999–1030.

[10] G. M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1203–1219.

[11] G. M. Lieberman, *Boundary regularity for solutions of degenerate parabolic equations*, Nonlinear Anal., 14 (1990), pp. 501–524.

[12] T. Namba, *Density-dependent dispersal and spatial distribution of a population*, J. Theoret. Biol., 86 (1980), pp. 351–363.

[13] S. I. Pohozaev, *On an approach to nonlinear equations*, Soviet Math. Dokl., 20 (1979), pp. 912–916. Translation of Dokl. Akad. Nauk. SSSR, 247 (1979), pp. 1327–1331.

[14] S. I. Pohozaev, *On a constructive method in the calculus of variations*, Soviet Math. Dokl., 37 (1988), pp. 274–277. Translation of Dokl. Akad. Nauk. SSSR, 298 (1988), pp. 1330–1333.

[15] S. I. Pohozaev, *On the method of fibering a solution in nonlinear boundary value problems*, Proc. Steklov Inst. Math., 3 (1992), pp. 157–173.

[16] P. Tolksdorf, *Regularity for a more general class of quasilinear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.

SIAM J. Math. Anal.
Vol. 31, No. 1, pp. 134–153

BOUNDS ON THE DISPERSION OF VORTICITY IN 2D INCOMPRESSIBLE, INVISCID FLOWS WITH A PRIORI UNBOUNDED VELOCITY[*]

J. HOUNIE[†], M. C. LOPES FILHO[‡], and H. J. NUSSENZVEIG LOPES[‡]

Abstract. We consider approximate solution sequences of the 2D incompressible Euler equations obtained by mollifying compactly supported initial vorticities in L^p, $1 \leq p \leq 2$, or bounded measures in H^{-1}_{loc} and exactly solving the equations. For these solution sequences we obtain uniform estimates on the evolution of the mass of vorticity and on the measure of the support of vorticity outside a ball of radius R. If the initial vorticity is in L^p, $1 \leq p \leq 2$, these uniform estimates imply certain a priori estimates for weak solutions which are weak limits of these approximations. In the case of nonnegative vorticities, we obtain results that extend, in a natural way, the cubic-root growth of the diameter of the support of vorticity proved first by C. Marchioro for bounded initial vorticities [*Comm. Math. Phys.*, 164 (1994), pp. 507–524] and extended by two of the authors to initial vorticities in L^p, $p > 2$.

Key words. incompressible flow, ideal flow, vorticity, irregular transport

AMS subject classifications. 35Q35, 76C05

PII. S0036141098337503

Introduction. The main object of this work is the behavior of weak solutions of the 2D Euler equations, modeling the flow of incompressible, inviscid ideal fluids in two space dimensions. We will be concerned with flows of fluids that are assumed to fully occupy the 2D Euclidean plane, with velocity vanishing at infinity. We write the initial value problem in the form of the *vorticity equation*:

$$\begin{cases} \omega_t + u \cdot \nabla \omega = 0 \ \text{ in } \mathbb{R}^2 \times (0,\infty), \\ \text{div } u = 0 \ \text{ in } \mathbb{R}^2 \times [0,\infty), \\ \text{curl } u = \omega \ \text{ in } \mathbb{R}^2 \times [0,\infty), \\ \omega(x,0) = \omega_0(x) \ \text{ on } \mathbb{R}^2 \times \{t=0\}. \end{cases} \tag{0.1}$$

The velocity can be eliminated from the vorticity equation by means of the Biot–Savart law:

$$u(x,t) = (K * \omega(\cdot,t))(x) \equiv \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{(x-y)^\perp}{|x-y|^2} \omega(y,t) dy.$$

The usual strategy to obtain existence of weak solutions to the problem (0.1) is to consider a suitable approximate problem, for which existence of solutions is known, and then to obtain enough estimates to pass to the limit in the weak form of the equations. The standard approximation schemes used in the literature are the following: smoothing out initial data, the vanishing viscosity limit of the Navier–Stokes equations, and desingularized vortex methods. In this work we are specifically

[*] Received by the editors April 16, 1998; accepted for publication (in revised form) February 5, 1999; published electronically November 23, 1999.

http://www.siam.org/journals/sima/31-1/33750.html

[†] Departamento de Matemática, UFSCar, São Carlos, SP 13565-905 Brazil (hounie@power.mat.ufscar.br). The research of this author was supported in part by CNPq grant 303.766/77-8.

[‡] Departamento de Matemática, IMECC-UNICAMP, Caixa Postal 6065, Campinas, SP 13081-970 Brazil (mlopes@ime.unicamp.br, hlopes@ime.unicamp.br). The research of M. C. Lopes Filho was supported in part by CNPq grant 300.962/91-6 and FAPESP grant 96/7635-4. The research of H. J. Nussenzveig Lopes was supported in part by CNPq grant 300.158/93-9.

concerned with weak solutions obtained by exactly solving (0.1) with smoothed-out initial data. If the initial vorticity ω_0 is a function in $L^p(\mathbb{R}^2)$, $1 < p < \infty$, with compact support, the existence of a weak solution obtained as the weak limit of a sequence of approximate solutions (produced by mollifying initial data) was first proved by DiPerna and Majda in [4]. For nonnegative initial vorticities in the space of bounded Radon measures with compact support, $\mathcal{BM}_c(\mathbb{R}^2)$, and in $H^{-1}_{\mathrm{loc}}(\mathbb{R}^2)$ a corresponding existence result was proved by Delort in [2]. Vecchi and Wu in [13] extended Delort's proof to initial vorticities of compact support in $L^1(\mathbb{R}^2) \cap H^{-1}_{\mathrm{loc}}(\mathbb{R}^2)$, without sign restrictions. Uniqueness, in these cases, is an outstanding open problem, as is existence for arbitrary bounded Radon measures of compact support in $H^{-1}_{\mathrm{loc}}(\mathbb{R}^2)$. Following DiPerna and Majda, we will refer to initial vorticities in $\mathcal{BM}(\mathbb{R}^2) \cap H^{-1}_{\mathrm{loc}}(\mathbb{R}^2)$ as vortex sheet initial data, which we will abbreviate with the acronym VSID. For bounded initial vorticities, then both existence and uniqueness of weak solutions were obtained by Yudovich in [14].

Little is known regarding the qualitative behavior of weak solutions of (0.1). The general problem we will focus on is the following: How fast can a fluid particle be displaced from its initial position and how is this displacement affected by the regularity of the subjacent flow? If the initial vorticity lies in the space $L^p_c(\mathbb{R}^2)$ (the space of compactly supported functions in L^p), $p > 2$, it is well known that the corresponding velocity field is bounded a priori. This means that the trajectory of almost all fluid particles is contained in a space-time cone centered at their initial positions and with aperture bounded by global conserved quantities of the flow. Since vorticity is constant along particle trajectories, this implies that the support of vorticity remains compact and its diameter grows at most linearly in time. For nonnegative bounded vorticities Marchioro [10] showed that the growth of the displacement from the initial position is at most of the order of the cubic-root of time, so that the space-time cone above can be substituted with a space-time cubic parabola. This result captures the trend that flows with single-signed vorticity have of rotating, rather than spreading particles. The result was extended by two of the authors in [9] to nonnegative initial vorticities in $L^p_c(\mathbb{R}^2)$, $p > 2$. However, the estimate on the aperture of the cubic parabola obtained is lost when $p \to 2^+$.

If the initial vorticity is in $L^p_c(\mathbb{R}^2)$, $1 \leq p \leq 2$, or in $\mathcal{BM}_c(\mathbb{R}^2) \cap H^{-1}_{\mathrm{loc}}(\mathbb{R}^2)$, it is not known whether the flow preserves the compactness of the support of vorticity. This problem was the initial motivation for the present work. The results we obtain here address the rate of dispersion of vorticity (or, equivalently, of material domains) in time. We will prove that the pictures obtained for more regular flows, i.e., linear cones in space-time for general vorticities and cubic parabolas for nonnegative vorticities, remain substantially true for even the most irregular cases. More precisely, we will show that, for any approximate solution sequence, given an initial disk in the plane and any $\varepsilon > 0$, there exists an aperture for a space-time cone (and for a cubic parabola in the case of nonnegative vorticity), uniform in the sequence, for which the set of particles in the initial disk whose trajectories leave the cone (respectively, the cubic parabola) has Lebesgue measure less than ε.

The remainder of this paper is organized in three sections: the first on flows without sign restriction on the vorticity, the second on flows with nonnegative vorticity, and the third containing extensions and conclusions. In the first one, we obtain estimates resembling Chebyshev inequalities for the Lagrangian maps that are applicable to any linear transport equation with a divergence-free smooth velocity field bounded in $L^q(\mathbb{R}^2)$. These results can be better understood in the context of the transport theory by vector fields with Sobolev space regularity by DiPerna and Lions [3]. In the specific context of the 2D vorticity equation, we also obtain a result of the same nature in the physically relevant situation where the velocity is only $L^2_{\mathrm{loc}}(\mathbb{R}^2)$. The second section begins with a simplified proof of an exponential decay estimate on the mass of vorticity near infinity due to Marchioro (this is the heart of the proof of Theorem 2.1 in [10]). We apply Marchioro's result and the Chebyshev inequalities obtained in the first section to get results on the smallness of the mass and of the Lebesgue measure of the support of vorticity outside a suitable cubic parabola. All our results are proved for a smooth approximate solution sequence generated by regularizing initial data, with estimates independent of the regularization parameter.

Some remarks regarding notation are in order. We denote by $B(p;R)$ the open ball centered at p with radius R in the plane. The Lebesgue measure of the set E is denoted by $|E|$ and the complement of E is denoted by E^c. If $z = (z_1, z_2)$ is a point in the plane, then $z^\perp = (-z_2, z_1)$. We denote the Lebesgue conjugate exponent of p by $p' = p/(p-1)$. Finally, we will use supp ω to denote the support of the function ω.

1. Chebyshev inequalities. We begin with a result which applies to a general flow by a divergence-free, time-dependent vector field u. Consider a bounded, divergence-free, smooth vector field $u : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$.

Let $X = X(\alpha, t)$ denote an orbit associated with the flow by u:

$$\begin{cases} \dfrac{dX}{dt} = u(X,t), \ 0 < t < T, \\ \\ X(\alpha, 0) = \alpha \in \mathbb{R}^n. \end{cases}$$

We use $X(D,t) \equiv \{X(\alpha,t) \mid \alpha \in D\}$ to denote the flow of a set $D \subseteq \mathbb{R}^n$ under the vector field u. We often refer to the family of diffeomorphisms $\alpha \mapsto X(\alpha,t)$ as Lagrangian maps.

The first result of this section will be referred to as the *filtering theorem.*

THEOREM 1.1. *Let* $0 < R_1 < R_2$ *and define the annulus* $A = \{x \in \mathbb{R}^n \mid R_1 < |x| < R_2\}$, $\Sigma(R_1, R_2, t) \equiv \{\alpha \in B(0;R_1) \mid |X(\alpha,t)| > R_2\}$ *and let* $q \geq 1$. *Then*

$$|\Sigma(R_1,R_2,t)| \leq \left(\frac{t \sup\limits_{0 \leq t \leq T} \|u(\cdot,t)\|_{L^q(A)}}{R_2 - R_1} \right)^q .$$

Proof. Fix $t > 0$. In this proof we will abbreviate $\Sigma(R_1, R_2, t)$ by Σ. Let us introduce the material cylinder $\mathcal{C}$ defined by

$$\mathcal{C} \equiv \bigcup_{0 \leq s \leq t} X(\Sigma, s).$$

The proof consists of integrating and estimating the radial component of velocity on the set $(A \times [0,t]) \cap \mathcal{C}$. Let $\chi_A = \chi_A(x)$ denote the characteristic function of the annulus A. Then, by incompressibility we have

$$\int_{\mathcal{C}} \chi_A(x) u(x,s) \cdot \frac{x}{|x|} dx ds = \int_0^t \int_\Sigma \chi_A(X(\alpha,s)) \frac{d|X(\alpha,s)|}{ds} d\alpha ds.$$

Claim. For any $\alpha \in \Sigma$, we have

$$\int_0^t \chi_A(X(\alpha,s)) \frac{d|X(\alpha,s)|}{ds} ds = R_2 - R_1. \tag{1.1}$$

To see that, consider $\Gamma \equiv \{0 < s < t \mid X(\alpha, s) \in A\}$. Since Γ is open, it can be written as a countable union of disjoint open intervals:

$$\Gamma = \bigcup_{i=1}^{\infty} (a_i, b_i).$$

Therefore,

$$\int_0^t \chi_A(X(\alpha,s)) \frac{d|X(\alpha,s)|}{ds} ds = \sum_{i=1}^{\infty} \int_{a_i}^{b_i} \frac{d|X(\alpha,s)|}{ds} ds$$

$$= \sum_{i=1}^{\infty} \left(|X(\alpha,b_i)| - |X(\alpha,a_i)|\right).$$

By the continuity of the trajectories $X(\alpha,\cdot)$, each of these numbers $|X(\alpha,b_i)|$ and $|X(\alpha,a_i)|$ is either R_1 or R_2. The curve $s \mapsto X(\alpha,s)$ has finite total length, and hence the summation above has a finite number of nonzero terms, which correspond precisely to the time intervals during which the curve completely traverses the annulus. Since $|X(\alpha,0)| = |\alpha| < R_1$ and $|X(\alpha,t)| > R_2$,

$$\sum_{i=1}^{\infty} \left(|X(\alpha,b_i)| - |X(\alpha,a_i)|\right) = R_2 - R_1,$$

and the claim is proved. $\square$

Hence, in view of (1.1),

$$\int_{\mathcal{C}} \chi_A(x) u(x,s) \cdot \frac{x}{|x|} dx ds = (R_2 - R_1)|\Sigma|. \tag{1.2}$$

On the other hand, we also have

$$\int_{\mathcal{C}} \chi_A(x) u(x,s) \cdot \frac{x}{|x|} dx ds \leq \int_0^t \int_{X(\Sigma,s)} |\chi_A(x) u(x,s)| dx ds$$

$$\leq \int_0^t \|\chi_A(\cdot) u(\cdot,s)\|_{L^q(X(\Sigma,s))} |\Sigma|^{(q-1)/q} ds$$

$$\leq t \sup_{0 \leq s \leq t} \|u(\cdot,s)\|_{L^q(A)} |\Sigma|^{(q-1)/q}.$$

Putting together the identity (1.2) and the inequality above we obtain the estimate we wished. $\square$

This result can be understood in the context of the linear transport theory developed by DiPerna and Lions in [3]. What we achieve is control over the local transport in terms of weak local control over the transporting vector fields that can be applied in situations where the flow is very singular. A theorem of this nature can also be proved for vector fields with bounded divergence, which is the context of [3]. However, in this work we are interested in the incompressible situation.

In [3], DiPerna and Lions observed that if the vector field $u \in L^q(\mathbb{R}^n) \cap W^{1,1}_{\text{loc}}(\mathbb{R}^n)$ has bounded divergence, then the Lagrangian maps are $L^q_{\text{loc}}(\mathbb{R}^n)$. Let $X(\cdot, t)$ be the unique renormalized flow associated with u, with $X(\cdot, t) \in L^q_{\text{loc}}(\mathbb{R}^n)$. In order to compare the estimate in Theorem 1.1 with results obtained by DiPerna and Lions, first recall the classical Chebyshev inequality, which states that if $\Omega \subset \mathbb{R}^n$ and if f in $L^q(\Omega)$, then for any $\lambda > 0$,

$$|\{x \in \Omega : |f(x)| > \lambda\}| \leq \frac{\|f\|^q_{L^q(\Omega)}}{\lambda^q}.$$

Note that we have

$$|\{\alpha \in B(0; R_1) \mid |X(\alpha, t)| > R_2\}| \leq |\{\alpha \in B(0; R_1) \mid |X(\alpha, t) - \alpha| > R_2 - R_1\}|$$

$$\leq \frac{\|X(\alpha, t) - \alpha\|^q_{L^q(B(0;R_1))}}{(R_2 - R_1)^q} = (1.1),$$

where the last inequality follows from the Chebyshev inequality applied to $X(\alpha, t) - \alpha \in L^q(B(0, R_1))$,

$$(1.1) = \frac{\| \int_0^t u(X(\alpha, s), s) ds\|^q_{L^q(B(0;R_1))}}{(R_2 - R_1)^q} \leq \frac{t^q \sup_{0 \leq s \leq t} \|u(\cdot, s)\|^q_{L^q(\mathbb{R}^n)}}{(R_2 - R_1)^q},$$

where the final inequality was deduced from the generalized Minkowski inequality; see [12]. The estimate in Theorem 1.1 is a generalization of this conclusion mainly because our estimate is local, in the sense that it depends only on the L^q-norm of u in the annulus A and not on a global L^q bound.

In the next result we single out a special case of the filtering theorem, which is more in the nature of a Chebyshev inequality for the Lagrangian maps, and which will be useful in the applications to 2D incompressible flow. Once again, we assume that the flow u is smooth.

Corollary 1.2. *Let $S_0 \subseteq B(0; R_0)$ and $t > 0$. Then, for every $R > R_0$, we have*

$$|X(S_0, t) \cap B(0; R)^c| \leq \left(\frac{t \sup\limits_{0 \leq s \leq t} \|u(\cdot, s)\|_{L^q(\mathbb{R}^n)}}{R - R_0} \right)^q.$$

Proof. Since

$$|X(S_0, t) \cap B(0; R)^c| = |\{X(\alpha, t) \mid \alpha \in S_0 \text{ and } |X(\alpha, t)| > R\}|,$$

we have, by incompressibility, that this is equal to

$$|\{\alpha \in S_0 \mid |X(\alpha, t)| > R\}| \leq |\Sigma(R_0, R, t)|,$$

and the result follows from Theorem 1.1. □

This estimate applies to incompressible flows of ideal fluids in a number of instances. First, the case $q = 2$ applies to n-dimensional incompressible flows as long as the flow exists and the initial data has globally bounded kinetic energy. In the remainder of this article we will develop applications to 2D Euler flows.

For any $1 \leq p \leq \infty$, the L^p-norm of vorticity is a conserved quantity as long as the flow is smooth. We first assume that $\omega_0 \in L^p_c$, and we are interested in the cases $1 < p \leq 2$. Our concern is the propagation of the support of vorticity, which is a material domain. In order to apply Corollary 1.2, we need to know the appropriate a priori estimate for velocity. This is given in the next result, which is an analogue of the Sobolev embedding $W^{1,p} \hookrightarrow L^{p^*}$, with $p^* = 2p/(2-p)$.

PROPOSITION 1.3. *If $\omega \in L^p(\mathbb{R}^2)$, for some $1 < p < 2$, then $u = K * \omega \in L^{p^*}(\mathbb{R}^2)$, where p^* is the critical Sobolev exponent introduced above. Moreover, we have the estimate*

$$\|u\|_{L^{p^*}} \leq \frac{C}{\sqrt{2-p}} \|\omega\|_{L^p},$$

for some $C = C(p)$, which blows up as $p \to 1$ and remains bounded as $p \to 2$.

Proof. Let I_1 be the first-order Riesz potential, so that, for f in the Schwarz space $\mathcal{S}(\mathbb{R}^n)$

$$\widehat{(I_1 f)}(\xi) = \frac{\widehat{f}(\xi)}{2\pi|\xi|}.$$

We consider also the Riesz transforms R_j in $\mathbb{R}^2$, $j = 1, 2$ so that, for a function $f \in \mathcal{S}(\mathbb{R}^n)$

$$\widehat{(R_j f)}(\xi) = i \frac{\xi_j \widehat{f}(\xi)}{|\xi|},$$

where $i = \sqrt{-1}$ and $\xi = (\xi_1, \xi_2)$.

The Riesz transforms are bounded in $L^q(\mathbb{R}^2)$, for any $1 < q < \infty$ with the operator norm continuous with respect to q, blowing up as $q \to 1$; see [12]. By the Hardy–Littlewood–Sobolev theorem, the Riesz potential maps $L^p(\mathbb{R}^2)$ continuously into $L^{p^*}(\mathbb{R}^2)$.

We observe that the Biot–Savart law can be rewritten (up to a constant factor) as

$$u = I_1 R^{\perp} \omega,$$

where $R^{\perp} = (-R_2, R_1)$.

To see this, we first note that $I_1 R^{\perp}$ maps L^p continuously into L^{p^*}. For a function in $f \in \mathcal{S}(\mathbb{R}^2)$ we have that $I_1 R^{\perp} f = K_1 * f$, where $K_1 \in \mathcal{S}'(\mathbb{R}^2)$ is such that its Fourier transform is $\widehat{K_1} = i\xi^{\perp}/(2\pi|\xi|^2)$, for $\xi \neq 0$.

Let ω be a vorticity in the Schwarz space $\mathcal{S}(\mathbb{R}^2)$, and consider both $u_1 = K_1 * \omega$ and $u_2 = K * \omega$. We will show they are the same. Observe that u_1 and u_2 are tempered distributions. The vector field u_2 is the unique solution to the elliptic system:

$$\begin{cases} \operatorname{div} u = 0, \\ \operatorname{curl} u = \omega, \\ |u| \to 0 \text{ as } |x| \to \infty. \end{cases}$$

We can pass the Fourier transform on the system above, and invert the resulting linear system for $\xi \neq 0$, to find that the Fourier transforms of u_1 and u_2 coincide. In particular, by varying ω, one may conclude that $\widehat{K} = \widehat{K_1}$, for $\xi \neq 0$, which then implies, by Theorem 3.2.3 of [5] (since $\widehat{K}$ and $\widehat{K_1}$ are homogeneous of degree $-1 > -2$), that $K_1 = K$ and hence that $u_1 = u_2$.

The proposition is proved, except for the asymptotic behavior, as $p \to 2$ of the operator norm of I_1.

To prove the asymptotic estimate, we begin by following the proof of Proposition 3.1.2 in [1], which gives a pointwise estimate of $I_1 f$ for $f \in L^p(\mathbb{R}^2)$ in terms of the maximal function $M|f|$. Tracking the constants, one arrives at the following estimate:

$$|I_1 f|(x) \leq C \left[2\pi + (2\pi)^{(p-1)/p} \left(\frac{p-1}{2-p} \right)^{(p-1)/p} \right] \|f\|_{L^p(\mathbb{R}^2)}^{p/2} (M|f|(x))^{p/p^*}.$$

Using the Hardy–Littlewood maximal theorem, this implies that

$$\|I_1 f\|_{L^{p^*}(\mathbb{R}^2)} \leq A_p \|f\|_{L^p(\mathbb{R}^2)},$$

with $A_p = C[2\pi + (2\pi)^{(p-1)/p}(\frac{p-1}{2-p})^{(p-1)/p}]$. Since $1 < p < 2$, A_p can be bounded from above by $C/\sqrt{2-p}$. □

Estimates for the propagation of support of vorticity can now be proved as a further corollary of Theorem 1.1 and Proposition 1.3.

COROLLARY 1.4. *Assume that ω_0 is a smooth function such that* supp $(\omega_0) \subseteq B(0; R_0)$. *If $1 < p < 2$, then there exists $C_p > 0$ such that, for any $R > R_0$,*

$$|\text{supp } \omega(\cdot, t) \cap B(0; R)^c| \leq \left(\frac{C_p \, t \, \|\omega_0\|_{L^p}}{R - R_0} \right)^{p^*}.$$

In addition, there exist constants $C > 0$ and $\eta > 0$ such that if $t/(R - R_0) < \eta$, then

$$|\text{supp } \omega(\cdot, t) \cap B(0; R)^c| \leq |\text{supp } \omega_0| \exp \left(-C \frac{(R - R_0)^2}{t^2 \|\omega_0\|_{L^2}^2} \right).$$

Proof. The first part is a trivial consequence of Corollary 1.2 and Proposition 1.3 together with the fact that the support of vorticity is a material domain: supp $\omega(\cdot, t) = X(\text{supp } \omega_0, t)$.

We now consider the second part. We use the fact that a compactly supported function in L^2 is also in L^p for any $p < 2$. More precisely, we have

$$\|\omega_0\|_{L^p} \leq |\text{supp } \omega_0|^{1/p^*} \|\omega_0\|_{L^2}.$$

Hence, from Proposition 1.3, we know that

$$|\text{supp } \omega(\cdot, t) \cap B(0; R)^c| \leq |\text{supp } \omega_0| \left(\frac{Ct\|\omega_0\|_{L^2}}{R - R_0} \right)^{p^*},$$

for any $1 < p < 2$. We optimize the estimate above in p. Since we are interested only in the behavior for p near 2, we restrict ourselves to searching for minima in the range $7/4 < p < 2$. We find that if

$$\frac{t}{R - R_0} < \frac{1}{2C\|\omega_0\|_{L^2}} e^{-7/16} \equiv \eta,$$

then there exists another constant $\widetilde{C} > 0$ such that

$$|\text{supp } \omega(\cdot, t) \cap B(0; R)^c| \leq |\text{supp } \omega_0| \exp \left(-\widetilde{C} \frac{(R - R_0)^2}{t^2 \|\omega_0\|_{L^2}^2} \right)$$

as we wanted. □

The critical estimate in terms of the L^2-norm obtained above can be understood as a Trudinger–Moser inequality for the Lagrangian maps. The proof we presented is a variation on the standard proofs in this context.

If the initial vorticity has compact support and it belongs to $L^1(\mathbb{R}^2)$ or it is a bounded Radon measure in $H^{-1}_{\text{loc}}(\mathbb{R}^2)$, that is, VSID, then the associated velocity belongs to $L^2_{\text{loc}}(\mathbb{R}^2)$ for each fixed time. It is well known that velocity belongs to $L^2(\mathbb{R}^2)$ only if the vorticity has vanishing integral over all of $\mathbb{R}^2$. Consequently, for initial vorticities of compact support in $L^1(\mathbb{R}^2)$ or VSID, with integral zero, we also have an estimate, valid for smooth approximate solution sequences, of the form

$$|\text{supp } \omega(\cdot,t) \cap B(0;R)^c| \leq \left(\frac{t\|u_0\|_{L^2}}{R-R_0}\right)^2.$$

However, flows with locally bounded kinetic energy are of physical interest. Furthermore, the only rigorous existence result for weak solutions with VSID requires that the initial vorticities have a distinguished sign. We can still prove an estimate for the Lebesgue measure of the support of vorticity lying outside a ball of radius R in this setting.

THEOREM 1.5. *Let ω_0 be a smooth function and let $T > 0$. Assume the support of ω_0 is contained in the ball $B(0;R_0)$. Then there exists a constant $C = C(T,R_0) > 0$ such that for all $R > R_0$ and $0 \leq t \leq T$ we have*

$$|\text{supp } \omega(\cdot,t) \cap B(0;R)^c| \leq C\left(\frac{t}{R-R_0}\right)^2.$$

Proof. Fix $R > R_0$ and $0 \leq t \leq T$. Recall

$$\Sigma = \Sigma(R_0,R,t) \equiv \{\alpha \in B(0;R_0) \mid |X(\alpha,t)| > R\}.$$

Observe that

$$|\text{supp } \omega(\cdot,t) \cap B(0;R)^c| \leq |\Sigma|.$$

Hence, it is enough to estimate $|\Sigma|$. We have

$$|\Sigma| \leq \frac{1}{R^2}\int_\Sigma |X(\alpha,t)|^2 d\alpha$$

$$\leq \frac{2}{R^2}\left(\int_\Sigma |X(\alpha,t)-\alpha|^2 d\alpha + \int_\Sigma |\alpha|^2 d\alpha\right)$$

$$\equiv \frac{2}{R^2}(\mathcal{I}_1 + \mathcal{I}_2).$$

Note that $X(\alpha,t) - \alpha = \int_0^t u(X(\alpha,s),s)ds$. We will need to make use of the DiPerna–Majda decomposition of an L^2_{loc} velocity (see [4]). To do this, choose a circularly symmetric, smooth, and compactly supported function $\bar{\omega} = \bar{\omega}(|x|)$ such that $\int_{\mathbb{R}^2} \bar{\omega}(|x|)dx = \int_{\mathbb{R}^2} \omega_0(x)dx$. Let $\bar{u} \equiv K * \bar{\omega}(|\cdot|)$. The stationary velocity field $\bar{u}$ is smooth and decays as $1/|x|$, as $|x| \to \infty$. Let $\check{\omega}(x,s) \equiv \omega(x,s) - \bar{\omega}(|x|)$ and $\check{u} \equiv K * \check{\omega}$. It was shown in [4] that

$$\|\check{u}\|^2_{L^2(\mathbb{R}^2)} \leq K(T) = \|\check{u}(\cdot,0)\|^2_{L^2(\mathbb{R}^2)} e^{cT}.$$

Let $\alpha \in \Sigma$. Using the decomposition $u = \bar{u} + \check{u}$ we have

$$|X(\alpha,t) - \alpha|^2 \leq \left(\int_0^t |\bar{u}(X(\alpha,s))| ds + \int_0^t |\check{u}(X(\alpha,s),s)| ds \right)^2$$

$$\leq \frac{Ct^2}{R^2} + Ct \int_0^t |\check{u}(X(\alpha,s),s)|^2 ds.$$

We can now estimate $\mathcal{I}_1$:

$$\mathcal{I}_1 \leq \frac{Ct^2}{R^2}|\Sigma| + Ct \int_0^t \int_\Sigma |\check{u}(X(\alpha,s),s)|^2 d\alpha ds$$

$$\leq \frac{Ct^2}{R^2} + Ct^2 K(T).$$

In order to estimate $\mathcal{I}_2$ note that if $\alpha \in \Sigma$, then

$$|X(\alpha,t) - \alpha| \geq R - R_0 \geq \frac{|\alpha|(R - R_0)}{R_0}.$$

Therefore,

$$|\alpha| \leq \frac{R_0}{R - R_0} |X(\alpha,t) - \alpha|.$$

We hence obtain

$$\mathcal{I}_2 \leq \left(\frac{R_0}{R - R_0} \right)^2 \mathcal{I}_1 \leq \left(\frac{R_0}{R - R_0} \right)^2 \left(\frac{Ct^2}{R^2} + Ct^2 K(T) \right).$$

Collecting these estimates, we finally get

$$|\Sigma| \leq \frac{2}{R^2} \left(\frac{Ct^2}{R^2} + Ct^2 K(T) \right) \left(1 + \left(\frac{R_0}{R - R_0} \right)^2 \right)$$

$$\leq \frac{Ct^2}{(R - R_0)^2},$$

since, at the same time, $R > R_0$ and $R > R - R_0$. □

There is one essential difference between this result and that of Corollary 1.4, which is the exponential growth of the constant C in T, while the constant in Corollary 1.4 did not depend on T.

For flows with no restrictions on the sign of vorticity, there is a well-known trend for paired eddies with vorticities of opposite sign to move off to infinity with constant speed. Since more than one such pair of eddies may be present in a given flow, with different average speeds, one expects that in some situations the diameter of the support of vorticity may grow linearly in time. We offer an explicit example from vortex dynamics to illustrate this behavior.

We consider the point-vortex evolution of four vortices, with initial configuration occupying the four vertices of the rectangle $[-a_0, a_0] \times [-b_0, b_0]$, with vorticity strength $+\omega$ at (a_0, b_0) and $(-a_0, -b_0)$ and with vorticity strength $-\omega$ at $(a_0, -b_0)$ and $(-a_0, b_0)$. This configuration is called a *vortex quadrupole*. The evolution preserves the quadrupole structure and is determined by a 2×2 system of ordinary differential equations for $(a(t), b(t))$, which is the position of the point vortex in the first quadrant. (In fact, by the reflexion method, the evolution of a vortex quadrupole is precisely the evolution of a single vortex in the first quadrant, regarded as a domain with boundary.) This 2×2 system is explicitly integrable, and the solution is given by the formulas

$$a(t) = \sqrt{\frac{1}{2}\left(q(t)^2 + 4k^2 + q(t)\sqrt{q(t)^2 + 4k^2}\right)},$$

$$b(t) = \frac{ka(t)}{\sqrt{a(t)^2 - k^2}},$$

where

$$c_0 = \sqrt{a_0^2 + b_0^2}, \quad k = a_0 b_0 / c_0, \quad q(t) = \frac{\omega t}{4\pi k} - \frac{b_0^2 - a_0^2}{c_0}.$$

From these formulas it can be seen that the diameter of the support of vorticity grows linearly in time.

Since this article was first distributed in preprint form, a continuous version of this example was obtained by Iftimie, Sideris, and Gamblin in [6].

2. Flows with vorticity of distinguished sign. In this section we will concentrate on 2D flows with nonnegative vorticity. Our objective is to derive results for flows with a priori unbounded velocity that capture the $\mathcal{O}(t^{1/3})$ growth on the diameter of the support of vorticity proved by Marchioro in [10] for bounded vorticities.

Our results rely heavily on an exponential decay estimate on the mass of vorticity far from the center of motion. Although originally proved for flows with bounded vorticities in [10], this estimate actually applies, with negligible changes in the original proof, to very singular vorticities such as weak solutions of (0.1) with VSID, obtained as limits of approximate solution sequences generated by regularizing initial data. This exponential decay estimate was derived by Marchioro in the course of proving Theorem 2.1 in [10] and has never been stated as an independent result. We will do so here and we will offer a simplified proof, in part for the sake of completeness, in which we avoid the use of dyadic decompositions. We note that an even simpler and more elegant proof of Marchioro's exponential decay estimate has been derived independently by Iftimie and Sideris [7].

We begin with an elementary technical lemma and then proceed to Marchioro's result.

Lemma 2.1. *Let $\phi = \phi(r) \geq 0$ be a function such that*

$$\int_0^\infty \phi(r) r^2 dr \equiv L < \infty.$$

Let $0 < \lambda < 1$ *and* $a > 0$. *Then*

$$\int_0^{\lambda a} \frac{r}{a(a-r)} \phi(r) r dr \leq \frac{L}{a^2(1-\lambda)^2}.$$

Proof. Set $F(r) = \int_r^{\lambda a} \phi(s) s ds$. Then

$$\int_0^{\lambda a} \frac{r}{a(a-r)} \phi(r) r dr = -\int_0^{\lambda a} \frac{r}{a(a-r)} F'(r) dr = \int_0^{\lambda a} \frac{1}{(a-r)^2} F(r) dr$$

$$\leq \frac{1}{a^2(1-\lambda)^2} \int_0^{\lambda a} F(r) dr \leq \frac{1}{a^2(1-\lambda)^2} \int_0^{\lambda a} \int_r^{\lambda a} \phi(s) s ds dr$$

$$= \frac{1}{a^2(1-\lambda)^2} \int_0^{\lambda a} \phi(s) s^2 ds \leq \frac{L}{a^2(1-\lambda)^2}$$

as we wanted. □

THEOREM 2.2 (see Marchioro [10]). *Let* ω_0 *be a smooth nonnegative function with support contained in* $B(0; R_0)$. *Let* $\omega = \omega(x,t)$ *be the unique smooth solution of the vorticity equation* (0.1) *with initial vorticity* ω_0. *For* $R > 0$ *define*

$$m_t(R) \equiv \int_{|x|>R} \omega(x,t) dx. \tag{2.1}$$

Then there exists a constant $C > 0$, *depending only on* $\int_{\mathbb{R}^2} \omega_0(x) dx$, *on the moment of inertia* $\int_{\mathbb{R}^2} |x|^2 \omega_0(x) dx$ *and on* R_0, *such that for any* $n \in \mathbb{N}$ *and any* $R > 0$ *satisfying* $nR_0 < R \leq (n+1)R_0$, *we have*

$$m_t(R + R_0) \leq \left(\frac{Ct}{(R-R_0)^3} \right)^n.$$

Proof. Let $W = W(r)$ be a nondecreasing smooth function such that $W(r) = 0$, if $r \leq R_0$ and $W(r) = 1$ if $r \geq 2R_0$. Let $R > R_0$. Set $\varphi = \varphi(y) \equiv W(|y| - (R - R_0))$ for $y \in \mathbb{R}^2$. Clearly, if $|y| > R_0 + R$ or $|y| < R$, φ is constant, and hence its first derivatives vanish. We will need the fact that the second derivatives of φ are uniformly bounded, independently of $R \geq R_0$. Indeed,

$$\frac{\partial^2 \varphi}{\partial y_i \partial y_j} = W''(|y| - (R - R_0)) \frac{y_i y_j}{|y|^2} + W'(|y| - (R - R_0)) \left(\frac{\delta_{ij}}{|y|} - \frac{y_i y_j}{|y|^3} \right).$$

Therefore, this second derivative vanishes outside $R < |y| < R + R_0$ and is bounded by $C(1 + 1/R) < C(1 + 1/R_0)$.

Following Marchioro, we introduce the smoothed-out version of $m_t(R)$:

$$\tilde{m}_t(R) = \int_{\mathbb{R}^2} \varphi(y) \omega(y,t) dy.$$

Then

$$\frac{d}{dt} \tilde{m}_t(R) = \int_{\mathbb{R}^2} \varphi(y) \omega_t(y,t) dy = \int_{\mathbb{R}^2} \nabla \varphi(y) u(y,t) \omega(y,t) dy$$

$$= \frac{1}{2}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} (\nabla\varphi(y) - \nabla\varphi(z))\, K(y-z)\omega(y,t)\omega(z,t)dydz.$$

We divide $\mathbb{R}^2$ into three regions: $O_1 = B(0, R/2)$, $O_2 = \{R/2 \le |x| < R\}$, and $O_3 = \{|x| \ge R\}$, and divide $\mathbb{R}^2 \times \mathbb{R}^2$ into the nine disjoint regions $O_{ij} = O_i \times O_j$. We first observe that

$$\int_{O_{ij}} (\nabla\varphi(y) - \nabla\varphi(z))\, K(y-z)\omega(y,t)\omega(z,t)dydz = 0$$

if both i and j are at most 2.

We begin by estimating the integral on O_{13}:

$$\frac{1}{2}\int_{O_{13}} (\nabla\varphi(y) - \nabla\varphi(z))\, K(y-z)\omega(y,t)\omega(z,t)dydz$$

$$= -\frac{1}{2}\int_{|y|<R/2}\int_{|z|\ge R} W'(|z| - (R-R_0))\frac{z}{|z|}K(y-z)\omega(y,t)\omega(z,t)dydz$$

$$\le C\left(\sup_{|z|\ge R}\left|\int_{|y|<R/2}\frac{z}{|z|}K(y-z)\omega(y,t)dy\right|\right)\int_{|z|\ge R}\omega(z,t)dz.$$

We will now make use of Lemma 2.1. Let $\phi(r,t) \equiv \int_0^{2\pi}\omega(r(\cos\theta,\sin\theta),t)d\theta$. Then,

$$\int_0^\infty \phi(r,t)r^2dr = \int_{\mathbb{R}^2}|x|\omega(x,t)dx \le \frac{1}{2}\int_{\mathbb{R}^2}(1+|x|^2)\omega(x,t)dx$$

$$= C\int_{\mathbb{R}^2}(1+|x|^2)\omega_0(x)dx \equiv L.$$

Note that, for $|z| \ge R$, we have

$$\left|\int_{|y|<R/2}\frac{z}{|z|}K(y-z)\omega(y,t)dy\right| \tag{2.2}$$

$$= \left|\int_{|y|<R/2}\frac{y}{|z|}\frac{(y-z)^\perp}{2\pi|y-z|^2}\omega(y,t)dy\right| \le \int_{|y|<|z|/2}\frac{|y|}{|z|(|z|-|y|)}\omega(y,t)dy$$

$$= \int_0^{|z|/2}\frac{r}{|z|(|z|-r)}\phi(r,t)rdr \le \frac{4L}{|z|^2} \tag{2.3}$$

using Lemma 2.1 with $a = |z|$ and $\lambda = 1/2$.

We conclude the estimate on O_{13} obtaining

$$\frac{1}{2}\int_{O_{13}} (\nabla\varphi(y) - \nabla\varphi(z))\, K(y-z)\omega(y,t)\omega(z,t)dydz \le C\frac{m_t(R)}{R^2}.$$

Similarly, on O_{31}

$$\frac{1}{2}\int_{O_{31}} \left(\nabla\varphi(y) - \nabla\varphi(z)\right) K(y-z)\omega(y,t)\omega(z,t)dydz \leq C\frac{m_t(R)}{R^2}.$$

Next observe that since the moment of inertia is conserved, $m_t(r) \leq C/r^2$. We now estimate the integral on $O_{23} \cup O_{33}$. We have

$$\frac{1}{2}\int_{O_{23}\cup O_{33}} \left(\nabla\varphi(y) - \nabla\varphi(z)\right) K(y-z)\omega(y,t)\omega(z,t)dydz$$

$$\leq C\left(\sup_{y,z\in\mathbb{R}^2} |(\nabla\varphi(y) - \nabla\varphi(z))K(y-z)|\right)\int_{|y|\geq R/2}\omega(y,t)dy\int_{|z|\geq R}\omega(z,t)dz$$

$$\leq C\frac{m_t(R)}{R^2},$$

similarly for $O_{32} \cup O_{33}$. Finally, observe that

$$\frac{1}{2}\int_{O_{33}} \left(\nabla\varphi(y) - \nabla\varphi(z)\right) K(y-z)\omega(y,t)\omega(z,t)dydz$$

$$\leq C(m_t(R))^2 \leq C\frac{m_t(R)}{R^2},$$

since the second derivatives of φ are bounded.

We have therefore shown that

$$\frac{d}{dt}\tilde{m}_t(R) \leq C\frac{m_t(R)}{R^2},$$

that is,

$$\tilde{m}_t(R) \leq \frac{C}{R^2}\int_0^t m_s(R)ds,$$

since $\tilde{m}_0(R) = 0$. Now we repeat the backwards induction argument of Marchioro.

Note that

$$\tilde{m}_t(R) \leq m_t(R) \leq \tilde{m}_t(R - R_0).$$

We now fix $n \in \mathbb{N}$ and R such that $nR_0 < R \leq (n+1)R_0$. By iterating backwards in time and in R we get

$$\tilde{m}_t(R) \leq \frac{C^n}{\prod_{i=0}^{n-1}(R - iR_0)^2}\int_0^t\int_0^{s_1}\dots\int_0^{s_{n-1}} m_{s_n}(R-(n-1)R_0)ds_n\dots ds_2 ds_1$$

$$\leq \frac{C^n t^n}{(n!)^3} \leq \frac{C^n t^n e^{3n}}{n^{3n}} \equiv \left(\frac{Ct}{n^3}\right)^n \leq \left(\frac{Ct}{(R-R_0)^3}\right)^n.$$

Since $\tilde{m}_t(R) \geq m_t(R + R_0)$, the conclusion follows. □

The result above offers no control on the mass of vorticity contained in the annulus $\{R_0 \leq |x| \leq 2R_0\}$. The proof could be modified, by suitably changing the definition of W, so that this absence of control would occur only on the annulus $\{R_0 \leq |x| \leq R_0 + \varepsilon\}$, with ε arbitrary. However, the constant C would blow up as $\varepsilon \to 0$. To obtain uniform control over the mass of vorticity outside a ball of radius $R_0 + \varepsilon$, we need to use the Chebyshev-type inequalities proved in the first section.

The following results are extensions of the statement of Theorem 2.1 in [10] to much more singular flows. We will continue using the notation $m_t(R)$ as in (2.1).

PROPOSITION 2.3. *Let ω_0 be a smooth nonnegative function, with support contained in $B(0; R_0)$. Let $1 < p \leq 2$ and $\|\omega_0\|_{L^p(\mathbb{R}^2)} \leq K$. Then, for every $\delta > 0$, there exists $b = b(K, \delta) > 0$ such that for any $t > 0$*

$$m_t((R_0^3 + bt)^{1/3}) < \delta.$$

Proof. Fix $0 < \delta < 1$. We start with the trivial observation that, for any $R > R_0$,

$$m_t(R) \leq \|\omega_0\|_{L^p} \left|\text{supp } \omega(\cdot, t) \cap B(0; R)^c\right|^{1/p'}.$$

From this and from Corollary 1.4 it follows that there exists $b_1 = b_1(K, \delta) > 0$ such that $m_t(R_0 + b_1 t) < \delta$ for all $t > 0$.

From Theorem 2.2, there exists $b_2 = b_2(K, \delta) > 0$ such that $m_t(2R_0 + (b_2 t)^{1/3}) < \delta$, again for every $t > 0$.

It is easy to see that one can choose $b = b(K, \delta)$ such that

$$\min\left\{R_0 + b_1 t, 2R_0 + (b_2 t)^{1/3}\right\} \leq (R_0^3 + bt)^{1/3},$$

and this concludes the proof. □

A similar result is still true in L^1; however, the constant b obtained does not depend uniformly on the L^1-norm of vorticity.

PROPOSITION 2.4. *Let $\{\omega_0^\varepsilon\}$ be a uniformly integrable family of nonnegative smooth functions, with support contained in $B(0; R_0)$. Then, for every $\delta > 0$, there exists $b = b(\delta)$ such that for any ε*

$$\int_{|x| > (R_0^3 + bt)^{1/3}} \omega^\varepsilon(x, t)\,dx < \delta.$$

Proof. Fix $\delta > 0$. By the definition of uniform integrability, there exists $\eta > 0$ such that for any $E \subseteq \mathbb{R}^2$, with $|E| < \eta$ and for any ε,

$$\int_E \omega_0^\varepsilon(\alpha)\,d\alpha < \delta.$$

Recall, from the proof of Theorem 1.5, that

$$|\Sigma(R_0, R, t)| = |\{\alpha \in B(0; R_0) \mid |X^\varepsilon(\alpha, t)| > R\}| \leq C \frac{t^2}{(R - R_0)^2},$$

where X^ε is the trajectory associated with the velocity field induced by ω^ε. Note from the proof of Theorem 1.5 that C does not depend on ε.

It is then possible to choose b_1 such that $|\Sigma(R_0, R_0 + b_1 t, t)| < \eta$. Next note that

$$\int_{|x| > R_0 + b_1 t} \omega^\varepsilon(x, t)\,dx = \int_{\Sigma(R_0, R_0 + b_1 t, t)} \omega_0^\varepsilon(\alpha)\,d\alpha < \delta.$$

The remainder of the argument follows precisely as in the proof of Proposition 2.3. □

Let $\omega_0 \in L^1_c(\mathbb{R}^2)$ be nonnegative. Consider any weak solution ω, obtained by mollifying the initial data ω_0, in such a way as to keep the support of the regularized vorticities inside $B(0;R_0)$. Such a weak solution was first shown to exist by Delort in [2]. Then Proposition 2.4 implies that, for any $\delta > 0$, there exists b such that $m_t((R_0^3 + bt)^{1/3}) < \delta$. Of course, Proposition 2.3 implies the same estimate for weak solutions obtained by regularizing initial vorticities in $L^p_c(\mathbb{R}^2)$. The subtle difference is that, for $1 < p \leq 2$, b depends on ω_0 through its L^p-norm. The dependence of b on ω_0 in the L^1 case is more delicate. (It depends on the modulus of continuity of ω_0, regarded as a measure.)

For ω_0 a nonnegative, compactly supported bounded measure in $H^1_{\text{loc}}(\mathbb{R}^2)$, we cannot prove a result of this nature for the approximate solution sequences obtained by mollifying ω_0. Theorem 2.2 remains valid in this situation, enabling the choice of b_2. However, we have no tools to choose b_1, i.e., to estimate the mass of vorticity outside $B(0;R)$, with R close to R_0. Proposition 2.4 cannot be used since, by the Dunford–Pettis theorem, regularizing ω_0 does not produce a uniformly integrable sequence. The best result we can obtain along these lines retains the asymptotic behavior as $t \to \infty$. It is a trivial consequence of Theorem 2.2 that, for every $\delta > 0$, there exists $b > 0$ such that

$$m_t(2R_0 + (bt)^{1/3}) < \delta.$$

In a sense, the results above, controlling the dispersion of the mass of vorticity, are unsatisfactory. We set out to study how much the fluid particles can get displaced by irregular fluid flow, and the control on the dispersion of the mass of vorticity, at first glance, does not give information in that respect. The next two results address this issue more precisely, demonstrating that the control on the dispersion of the mass of vorticity achieved so far, plus the techniques and results of the first section, do indeed control the dispersion of material domains in general. We cast the results in terms of the measure of the set of vorticity-bearing particles flung far from their initial positions by the flow, a very particular material domain, but this restriction is not essential. We prove two results: one showing precisely how the control on the dispersion of the mass of vorticity implies control on particle trajectories and the second one, giving a less precise, but more elegant description, in which we bring out explicitly the cubic parabola behavior of the dispersion discovered by Marchioro.

Theorem 2.5. *Let ω_0 be a nonnegative smooth function, with support contained in $B(0;R_0)$. Suppose that $\|\omega_0\|_{L^1(\mathbb{R}^2)} \leq K$. Then there exist $C_1 = C_1(K,R_0) > 0$ and $C_2 = C_2(K,R_0) > 0$ such that if $R > 2R_0$ and $0 < t < C_2R^3$, then*

$$|\text{supp }\omega(\cdot,t) \cap B(0;R)^c| \leq \frac{C_1 t R^3}{C_2R^3 - t} m_t(R/4).$$

Proof. Fix $R > 2R_0$ and $t > 0$. Let $\Sigma(R/2,R,t) \equiv \{\alpha \in B(0;R/2) \mid |X(\alpha,t)| > R\}$, which we abbreviate Σ_R. Then, $|\text{supp }\omega(\cdot,t) \cap B(0;R)^c| \leq |\Sigma_R|$.

We intend to estimate the velocity in the annulus $A_R \equiv \{R/2 < |x| < R\}$. We decompose the velocity $u = K * \omega(\cdot,s)$, for $0 \leq s \leq t$, into a near-field and a far-field velocity in the following way:

$$u^N(x,s) = \int_{|y|>R/4} K(x-y)\omega(y,s)dy; \quad u^F = u - u^N.$$

Consider the cylinder $\mathcal{C}$, defined by $\mathcal{C} \equiv \bigcup_{0\le s\le t} X(\Sigma_R, s)$. Next, observe that

$$\int_{\mathcal{C}} \chi_{A_R}(x)u^N(x,s)\cdot\frac{x}{|x|}dxds$$

$$= \int_{\mathcal{C}} \chi_{A_R}(x)u(x,s)\cdot\frac{x}{|x|}dxds - \int_{\mathcal{C}} \chi_{A_R}(x)u^F(x,s)\cdot\frac{x}{|x|}dxds$$

$$\geq (R - R/2)|\Sigma_R| - t|\Sigma_R| \sup_{(x,s)\in\mathcal{C}} \left| u^F(x,s)\cdot\frac{x}{|x|}\chi_{A_R}(x)\right|$$

$$\geq \frac{R}{2}|\Sigma_R| - t|\Sigma_R|\frac{C}{R^2} = \left(\frac{R}{2} - \frac{Ct}{R^2}\right)|\Sigma_R|,$$

where the latter inequality follows from the same reasoning as in (2.2)–(2.3), with $C = C(K, R_0)$, and the former inequality is a consequence of (1.1). On the other hand, using Hölder's inequality we have

$$\int_{\mathcal{C}} \chi_{A_R}(x)u^N(x,s)\cdot\frac{x}{|x|}dxds \leq t \sup_{0\le s\le t} \|u^N(\cdot,s)\|_{L^1(A_R)}$$

$$\leq \widetilde{C}t \sup_{0\le s\le t} \int_{|y|>R/4} \omega(y,s) \int_{A_R} \frac{1}{|x-y|}dxdy$$

$$\leq \widetilde{C}tRm_t(R/4),$$

since, due to the monotonicity of $1/r$, the integral of $|x|^{-1}$ on any set of measure $3\pi R^2/4$ is maximized by taking the integrating set to be the ball with this measure centered at 0, and hence it is bounded by $\sqrt{3}\pi R$. Taking $C_2 = (2C)^{-1}$, $C_1 = \widetilde{C}/C$, and assuming that $t < C_2R^3$ we obtain the desired conclusion. □

Proposition 2.6. *Let ω_0 be a nonnegative smooth function with support contained in $B(0;R_0)$. Suppose that $\|\omega_0\|_{L^1(\mathbb{R}^2)} \leq K$. Then for every $\delta > 0$, there exists $b = b(K, R_0, \delta) > 0$ such that for every $t > 0$*

$$|\text{supp } \omega(\cdot,t) \cap B(0;(R_0^3 + bt)^{1/3})^c| < \delta.$$

Proof. The proof follows the reasoning of the proofs of Propositions 2.3 and 2.4, but it is more intricate.

We begin by choosing $b_1 > 0$ such that for $0 \leq t \leq 1$

$$|\text{supp } \omega(\cdot,t) \cap B(0;R_0 + b_1t)^c| < \delta.$$

This is done using Theorem 1.5, with $T = 1$.

Next we choose $b_2 > 0$ such that for $t \geq 1$

$$|\text{supp } \omega(\cdot,t) \cap B(0;8R_0 + (b_2t)^{1/3})^c| < \delta.$$

This is accomplished using Theorems 2.5 and 2.2, as we shall describe below.

Denote $R(t) = 8R_0 + (b_2 t)^{1/3}$ for some b_2 to be determined.

First we choose b_2 large enough so as to guarantee that, for some $C_3 > 0$, we have $t < t + C_3 < C_2(R(t))^3$, where C_2 comes from Theorem 2.5. Next, we invoke the estimate on $m_t(R(t)/4)$, given by Theorem 2.2, together with Theorem 2.5. After a number of straightforward estimates, we observe it is enough to find b_2 large enough, so that

$$C_4 b_2 (1 + t^2) \left(\frac{C_5}{b_2} \right)^{C_6 (b_2 t)^{1/3}} < \delta$$

for certain constants C_4, C_5, and C_6. Then it is enough to note that if b_2 is large enough, the left-hand side of the inequality above is monotone decreasing as a function of t and its value at $t = 1$ converges to zero as $b_2 \to \infty$.

Finally, it is easy to see that we can choose b so that

$$\begin{aligned} R_0 + b_1 t &\leq (R_0^3 + bt)^{1/3}, && 0 \leq t \leq 1, \\ 8R_0 + (b_2 t)^{1/3} &\leq (R_0^3 + bt)^{1/3}, && 1 \leq t < \infty. \end{aligned}$$

This concludes the proof. □

3. Concluding remarks. We have proved several results concerning the amount of vorticity near infinity, arising from flow with compactly supported initial vorticity, explicitly formulated as uniform estimates on approximate solution sequences. Let us now consider a weak solution u of the incompressible 2D Euler equations, obtained as the weak limit of an approximate solution sequence u^ε. Let $\omega =$ curl u, $\omega^\varepsilon =$ curl u^ε. We restrict ourselves to approximate solution sequences obtained by mollifying the initial data and exactly solving the equations.

First we assume that $\omega_0 = \omega(\cdot, 0) \in L^p_c(\mathbb{R}^2)$, $1 \leq p \leq 2$. If $p = 1$, we have to assume in addition that $u_0 = u(\cdot, 0) \in L^2_{\text{loc}}(\mathbb{R}^2)$. The estimates obtained in Proposition 2.6 for nonnegative vorticities, in Corollary 1.4 for vorticities without sign restriction and $p > 1$, and in Theorem 1.5 if $p = 1$, which are all estimates on the size of the support of vorticity near infinity do not extend in any obvious way to the weak limit, because the measure of the support of an L^p function is not even weakly lower semicontinuous. These uniform estimates on the size of the support of vorticity imply estimates on the pth power integral of vorticity near infinity, which remain valid for the weak limit because the pth power integral is a weakly lower semicontinuous functional over L^p.

Let us be more precise. We will detail the argument in the case of a nonnegative initial vorticity $\omega_0 \in L^p_c(\mathbb{R}^2)$, $1 \leq p \leq 2$, supported in $B(0; R_0)$. We begin by observing that the mollified initial data ω_0^ε is uniformly pth power integrable, by the Dunford–Pettis theorem, since $|\omega_0^\varepsilon|^p$ converges strongly in L^1 to $|\omega_0|^p$. This means that, for every $\eta > 0$, there exists $\delta > 0$, independent of ε, such that

$$\int_E |\omega_0^\varepsilon|^p dx \leq \eta,$$

for any measurable set E with Lebesgue measure less than δ. Fix $\eta > 0$ and consider the corresponding δ. We use Proposition 2.6 to obtain $b > 0$, depending only on $\|\omega_0\|_{L^1}$, on R_0, and on δ such that

$$|\text{supp } \omega^\varepsilon(\cdot, t) \cap B(0; (R_0^3 + bt)^{1/3})^c| < \delta.$$

Hence, if E^ε is the backwards flow through u^ε of supp $\omega^\varepsilon(\cdot,t)\cap B(0;(R_0^3+bt)^{1/3})^c$, then

$$\int_{|x|>(R_0^3+bt)^{1/3}} |\omega^\varepsilon|^p(x,t)dx = \int_{E^\varepsilon} \omega_0^\varepsilon(x)dx \leq \eta,$$

since $|E^\varepsilon|<\delta$. Clearly, by the weak lower semicontinuity,

$$\int_{|x|>(R_0^3+bt)^{1/3}} |\omega|^p(x,t)dx \leq \eta.$$

Analogous results for vorticity without sign restrictions follow from Corollary 1.4 and Theorem 1.5 in the same manner.

For VSID, the only result obtained that extends, in the sense above, to an estimate on the weak limits is Theorem 2.2, again because the total variation of measures is weak-$*$ lower semicontinuous.

We have mentioned that there is no obvious way to pass the weak limit in the estimates on the size of the support of vorticity near infinity. We will show now that if $\omega_0\in L^p_c(\mathbb{R}^2)$, $1<p\leq 2$, then we can produce an estimate on the size of the support near infinity of any weak solution obtained as a weak limit of smooth approximants generated by mollifying initial data.

THEOREM 3.1. *Let $\omega_0\in L^p_c(\mathbb{R}^2)$, $1<p\leq 2$, and let u be a weak solution of the 2D Euler equations, obtained as the weak limit of an approximate solution sequence $\{u^\varepsilon\}$, obtained by mollifying the initial data. Let $\omega^\varepsilon=$ curl u^ε. Then, for almost every $t>0$ and every $R>0$*

$$|\text{supp }\omega(\cdot,t)\cap B(0;R)^c| \leq \limsup_{\varepsilon\to 0} |\text{supp }\omega^\varepsilon(\cdot,t)\cap B(0;R)^c|.$$

Proof. Fix $T>0$ and let $R_0>0$ be such that $B(0;R_0)$ contains the support of ω_0^ε for every ε.

We begin by observing that $\omega=\text{curl }u$ is the unique renormalized solution of the linear transport equation:

$$\begin{cases} f_t+u\cdot\nabla f=0, \\ f(x,0)=\omega_0, \end{cases} \tag{3.1}$$

as defined by DiPerna and Lions; see [3].

To see this, first note that the restriction to $p>1$ is needed to ensure that $u\in L^1([0,T];W^{1,1}_{\text{loc}}(\mathbb{R}^2))$. Of course, div $u=0$. We also have that

$$\frac{u}{1+|x|}\in L^1([0,T];L^2(\mathbb{R}^2))+L^1([0,T];L^\infty(\mathbb{R}^2)). \tag{3.2}$$

(See Remark 1.1 and Section 1.C of [4].) In order to show that ω is a renormalized solution of (3.1), we will make use of the uniqueness result, Theorem II.3, and the stability result, Theorem II.4, in [3]. Both of these results require u as above except that $u/(1+|x|)$ has to belong to $L^1([0,T];L^1(\mathbb{R}^2))+L^1([0,T];L^\infty(\mathbb{R}^2))$. However, it is easy to see that one can substitute this condition with (3.2) and still prove uniqueness and stability.

Next we check the hypothesis of the stability result. We know that ω is the weak-$*$ limit in $L^\infty([0,T];L^p(R^2))$ of the sequence $\{\omega^\varepsilon\}$, which is a smooth solution (and hence a renormalized solution) of (3.1) with u replaced by u^ε and ω_0 replaced by ω_0^ε. Additionally, the initial data ω_0^ε converge strongly to ω_0 in L^p. Therefore, by Theorem II.4 of [3], ω is the unique renormalized solution of (3.1).

Let X^ε be the Lagrangian map associated with the flow u^ε and X be the unique renormalized Lagrangian map associated with u, by Theorem III.2 of [3]. Then, by a time-dependent version of the stability of Lagrangian maps, Corollary III.1 in [3], we conclude that $X^\varepsilon \to X$ locally uniformly in time and locally in measure in space. Therefore, for every $\eta > 0$, there exists ε_0 independent of $t \in [0,T]$ such that for $\varepsilon < \varepsilon_0$,

$$|\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t) - X(\alpha,t)| > \eta\}| \leq \eta.$$

Fix $\eta > 0$ and choose $\varepsilon_0 = \varepsilon_0(\eta)$ as above. Let $R > 0$. Then, for any $\varepsilon < \varepsilon_0$, we have that

$$\begin{aligned} |\{\alpha \in B(0;R_0) | |X(\alpha,t)| > R\}| &\leq |\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t) - X(\alpha,t)| > \eta\}| \\ &\quad + |\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t)| > R-\eta\}| \\ &\leq \eta + |\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t)| > R-\eta\}| \\ \leq \eta + |\{\alpha \in B(0;R_0) | R-\eta < |X^\varepsilon(\alpha,t)| &\leq R\}| + |\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t)| > R\}| \\ &\leq \eta + (2\pi R\eta - \eta^2) + |\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t)| > R\}|, \end{aligned}$$

where the last inequality follows from the fact that X^ε is area-preserving. Therefore, taking $\limsup_{\eta\to 0}$, we have

$$|\{\alpha \in B(0;R_0) | |X(\alpha,t)| > R\}| \leq \limsup_{\varepsilon\to 0} |\{\alpha \in B(0;R_0) | |X^\varepsilon(\alpha,t)| > R\}|,$$

since we may assume that $\varepsilon_0(\eta) \to 0$ as $\eta \to 0$.

We conclude by observing that the renormalized Lagrangian map X is area-preserving and $\omega(X(\alpha,t),t) = \omega_0(\alpha)$ (see Theorem III.2 in [3]); hence

$$|\{\alpha \in B(0;R_0) | |X(\alpha,t)| > R\}| = |\text{supp } \omega(\cdot,t) \cap B(0;R)^c|. \qquad \square$$

Let u be a weak solution of the incompressible 2D Euler equations with vorticity $\omega = \text{curl } u \in L^\infty([0,T];L^p(\mathbb{R}^2))$, $p > 1$. It was mentioned in the proof of Theorem 4.1 in [8] that ω is the unique renormalized solution of the vorticity equation, regarded as a linear transport equation. We included an outline of the proof of this fact for the sake of completeness. Of course, given Theorem 3.1 above, the uniform estimates derived in Corollary 1.4 and in Proposition 2.6 remain valid for the weak solution, if the initial vorticity belongs to L^p_c, $p > 1$.

One natural question at this point is how close these estimates are to being optimal. In this respect, we do not have anything to add to the comments made by Marchioro in [10] and we refer the reader to the discussion contained there. We are left with no answer to the question we started with, i.e., whether the support of vorticity remains compact as time evolves. We can say only that the knowledge developed here does not appear to be enough to answer this question.

Finally, we note that since this article was first distributed in preprint form, a significant improvement of Marchioro's cubic-root estimate was obtained by Serfati; see [11]. A similar, if slightly weaker improvement was obtained independently by Iftimie and Sideris in [6]. By using the conservation of the center of vorticity, they observe that Marchioro's space-time cubic-root parabola can almost be improved to a fourth-root parabola. It would be possible to rewrite section 2 of our work reflecting these improvements, in a straightforward manner.

REFERENCES

[1] D. R. Adams and L. I. Hedberg, *Function Spaces and Potential Theory*, Grundlehren Math. Wiss. Vol. 314, Springer-Verlag, Berlin, 1996.

[2] J.-M. Delort, *Existence de nappes de tourbillon en dimension deux*, J. Amer. Math. Soc., 4 (1991), pp. 553–586.

[3] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[4] R. J. DiPerna and A. J. Majda, *Concentrations in regularizations for 2-D incompressible flow*, Comm. Pure Appl. Math., 40 (1987), pp. 301–345.

[5] L. Hörmander, *The Analysis of Linear Partial Differential Operators,* I, Grundlehren Math. Wiss. Vol. 256, Springer-Verlag, Berlin, 1983.

[6] D. Iftimie, T. Sideris, and P. Gamblin, *On the evolution of compactly supported planar vorticity*, Comm. Partial Differential Equations, 24 (1999), pp. 1709–1730.

[7] D. Iftimie and T. Sideris, private communication.

[8] P.-L. Lions, *Mathematical Topics in Fluid Mechanics Vol.* I: *Incompressible Models*, Oxford Lecture Series in Mathematics and Its Applications 3, Clarendon Press, Oxford University Press, New York, 1996.

[9] M. C. Lopes Filho and H. J. Nussenzveig Lopes, *An extension of C. Marchioro's bound on the growth of a vortex patch to flows with L^p vorticity*, SIAM J. Math. Anal., 29 (1998), pp. 596–599.

[10] C. Marchioro, *Bounds on the growth of the support of a vortex patch*, Comm. Math. Phys., 164 (1994), pp. 507–524.

[11] Ph. Serfati, *Borne en Temps des Caractéristiques de l'Équation d'Euler 2D à Tourbillon Positif et Localisation pour le Modèle Point-Vortex*, preprint, 1998.

[12] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[13] I. Vecchi and S. Wu, *On L^1-vorticity for 2-D incompressible flow*, Manuscripta Math., 78 (1993), pp. 403–412.

[14] V. I. Yudovich, *Non-stationary flow of an ideal incompressible liquid*, USSR Comp. Math. Math. Phys., 3 (1963), pp. 1407–1456.

SIAM J. MATH. ANAL.
Vol. 31, No. 1, pp. 154–165

A NECESSARY AND SUFFICIENT CONDITION FOR PALAIS–SMALE CONDITIONS[*]

KUAN-JU CHEN[†] AND HWAI-CHIUAN WANG[†]

Abstract. In this paper we prove the following assertions: (1) $\alpha_I = \alpha_M = \alpha_\Gamma = \alpha_{\Gamma'}$; (2) let $\Omega_0 = \Omega_1 \cup \Omega_2$, where $\Omega_1 \cap \Omega_2$ is bounded, and let $\alpha_i = \alpha(\Omega_i)$ be the index of J in Ω_i for $i = 0, 1, 2$. J satisfies the $(PS)_{\alpha_0}$-condition if and only if the inequality $\alpha_0 < \min\{\alpha_1, \alpha_2\}$ holds; (3) the union of a solvable domain and an unsolvable domain may be solvable and the union of two unsolvable domains may be solvable.

Key words. Palais–Smale condition, solvable domain

AMS subject classifications. 35J20, 35J25

PII. S0036141098338016

1. Introduction.

The study of Esteban–Lions domains is the cornerstone and the starting point for understanding the existence of solutions of equations in unbounded domains. In this paper, we examine the existence of solutions in perturbed Esteban–Lions domains.

Let $N \geq 2$ and $2 < p < 2^*$, where $2^* = \frac{2N}{N-2}$ for $N \geq 3$, $2^* = \infty$ for $N = 2$. Consider the semilinear elliptic equation

$$\begin{cases} -\Delta u + u = |u|^{p-2}u \quad \text{in} \quad \Omega, \\ \quad u \in H_0^1(\Omega), \end{cases} \tag{1.1}$$

where Ω is a domain in $\boldsymbol{R}^N$ and $H_0^1(\Omega)$ is the Sobolev space in Ω. It is well known that (1.1) in a bounded domain or in the whole space $\boldsymbol{R}^N$ admits a positive solution, but the same equation in an Esteban–Lions domain does not admit any solution. The Esteban–Lions domain is defined as follows.

DEFINITION 1.1. *We say that a proper unbounded domain Ω in $\boldsymbol{R}^N$ is an Esteban–Lions domain if there is $\chi \in \boldsymbol{R}^N$, $||\chi|| = 1$ such that $n(x)\cdot\chi \geq 0$, $n(x)\cdot\chi \not\equiv 0$ on $\partial\Omega$, where $n(x)$ denotes the unit outward normal to $\partial\Omega$ at the point x.*

Let the potential operators $a : H_0^1(\Omega) \to \boldsymbol{R}$, $b : H_0^1(\Omega) \to \boldsymbol{R}$, and the energy functional $J : H_0^1(\Omega) \to \boldsymbol{R}$ be given by

$$a(u) = \int_\Omega \left(|\nabla u|^2 + u^2\right),$$
$$b(u) = \int_\Omega |u|^p,$$
$$J(u) = \frac{1}{2}a(u) - \frac{1}{p}b(u).$$

In the following definitions, we simply denote Palais–Smale by (PS).

DEFINITION 1.2.

(1) *For $\beta \in \boldsymbol{R}$, a sequence $\{u_n\} \subset H_0^1(\Omega)$ is a $(PS)_\beta$-sequence for J if $J(u_n) \to \beta$ and $J'(u_n) \to 0$ strongly as $n \to \infty$;*

[*] Received by the editors April 9, 1998; accepted for publication (in revised form) December 17, 1998; published electronically December 3. 1999.
http://www.siam.org/journals/sima/31-1/33801.html

[†] Department of Mathematics, National Tsing Hua University, Hsinchu, 300, Taiwan, Republic of China (d837201@math.nthu.edu.tw, hwang@math.nthu.edu.tw).

(2) *$\beta \in \boldsymbol{R}$ is a $(PS)_\beta$-value for J if there is a $(PS)_\beta$-sequence for J;*

(3) *J satisfies the $(PS)_\beta$-condition if every $(PS)_\beta$-sequence for J contains a convergent subsequence;*

(4) *J satisfies the (PS) condition if, for every $\beta \in \boldsymbol{R}$, every $(PS)_\beta$-sequence for J contains a convergent subsequence.*

Note that J is of class $C^{1,1}$ (see Rabinowitz [9, Proposition B.10]) and J clearly satisfies the mountain pass hypothesis: there are r, $\delta > 0$ and $e \in H_0^1(\Omega)$, such that $e \notin \overline{B(0,r)}$, $J(e) = 0$, $J(u) \geq \delta > 0$ for $u \in \partial B(0,r)$.

Let

$$\begin{aligned}
&I = \inf\{a(u) \mid b(u) = 1\};\\
&\alpha_I = (\tfrac{1}{2} - \tfrac{1}{p}) I^{p/(p-2)};\\
&\mathbf{M} = \left\{u \in H_0^1(\Omega) \backslash \{0\} \mid a(u) = b(u)\right\};\\
&\alpha_M = \inf_{v \in M} J(v);\\
&\Gamma = \left\{g \in C([0,1], H_0^1(\Omega)) \mid g(0) = 0, g(1) = e\right\}, \text{ where } J(e) = 0;\\
&\alpha_\Gamma = \inf_{g \in \Gamma} \max_{t \in [0,1]} J(g(t));\\
&\Gamma' = \left\{K \subset H_0^1(\Omega) \,|K \text{ is closed, connected, and } 0,\ e \in K\right\};\\
&\alpha_{\Gamma'} = \inf_{K \in \Gamma'} \max_{u \in K} J(u).
\end{aligned}$$

Next we assert the following theorem.

Theorem A. $\alpha_I = \alpha_M = \alpha_\Gamma = \alpha_{\Gamma'}$.

For convenience we state the following definition.

Definition 1.3.

(1) *We say that $\alpha(\Omega) = \alpha_I$ is the index of the energy functional J in Ω;*

(2) *We say that a solution u of equation* (1.1) *is a ground state solution if $J(u) = \alpha(\Omega)$, and is a higher energy solution if $J(u) > \alpha(\Omega)$.*

Remark 1.

(1) *If the Nehari minimization problem α_M or the minimax problem α_Γ admits a solution u, then u must be a ground state solution. If J satisfies the $(PS)_{\alpha(\Omega)}$-condition, then the Nehari minimization problem α_M and the minimax problem α_Γ admit a ground state solution.*

(2) *Applying Theorem* A, *we prove that α_Γ is independent of the choice of e.*

(3) *Rabinowitz* [9, *p.* 19] *asked whether $\alpha_\Gamma = \alpha_{\Gamma'}$. We answer his question for our special energy functional J.*

(4) *Theorem* A *simplifies many calculations.*

Let $\Omega_0 = \Omega_1 \cup \Omega_2$, where $\Omega_1 \cap \Omega_2$ is bounded,

$$\mathbf{M}_i = \left\{u \in H_0^1(\Omega_i) \backslash \{0\} \mid a(u) = b(u)\right\},$$

and $\alpha_i = \alpha(\Omega_i)$ be the index of J in Ω_i for $i = 0, 1, 2$.

In this article, we will study the existence of positive solutions of (1.1) in a proper unbounded domain Ω. In fact, we give a necessary and sufficient condition in Ω, in which J satisfies the $(\mathrm{PS})_{\alpha(\Omega)}$-condition as follows.

Theorem B. *J satisfies the $(PS)_{\alpha_0}$-condition if and only if the inequality $\alpha_0 < \min\{\alpha_1, \alpha_2\}$ holds. In particular, if the inequality $\alpha_0 < \min\{\alpha_1, \alpha_2\}$ holds, then there is a solution u_0 of* (1.1) *in Ω_0.*

In section 4, Theorem B is applied to prove that the union of a solvable domain and an unsolvable domain may be solvable. We also assert that the union of two unsolvable domains may be solvable. Here we need another definition.

DEFINITION 1.4. *We say that Ω is solvable if there exists a positive solution of* (1.1) *in Ω; otherwise, Ω is unsolvable.*

2. Palais–Smale values. In this section, we study the set of all $(\mathrm{PS})_\beta$-values for J.

Let $\{u_n\} \subset H_0^1(\Omega)$ be a $(\mathrm{PS})_\beta$-sequence for J; then clearly $\beta \geq 0$ and $\{u_n\}$ is bounded in $H_0^1(\Omega)$. Let $\mathbf{F}$ be the set of all $(\mathrm{PS})_\beta$-values for J, where $\beta > 0$.

Moreover, the three important numbers α_I, α_M, and α_Γ are in $\mathbf{F}$.

LEMMA 2.1. *α_I, α_M, and α_Γ are (PS)-values for J.*

Proof. Lien–Tzeng–Wang [7] proved that α_I is a $(\mathrm{PS})_{\alpha_I}$-value for J. Using two different methods, the Ekeland variational principle and the deformation lemma, Brezis–Nirenberg [2] prove that α_Γ is a $(\mathrm{PS})_{\alpha_\Gamma}$-value for J.

Using the Ekeland variational principle, Stuart [11, Lemma 3.4] asserted that there is a $(\mathrm{PS})_{\alpha_M}$-sequence as well as a minimizing sequence for α_M in $H_0^1(\Omega)$. We generalize his result and prove that every minimizing sequence for α_M is a $(\mathrm{PS})_{\alpha_M}$-sequence for J as follows.

Let $\{u_n\} \subset \mathbf{M}$ be a minimizing sequence for $J : a(u_n) = b(u_n)$ for all $n = 1, 2, 3, \ldots$ and $J(u_n) = (\frac{1}{2} - \frac{1}{p})a(u_n) = \alpha_M + o(1)$ as $n \to \infty$. Then

$$a(u_n) = \frac{2p}{p-2}\alpha_M + o(1) \text{ as } n \to \infty. \tag{2.1}$$

For $n = 1, 2, \ldots$, denote

$$f_n(\varphi) = \int_\Omega |u_n|^{p-2} u_n \varphi, \quad \varphi \in H_0^1(\Omega).$$

Let $\phi \in H_0^1(\Omega)$ and $||\phi||_{H^1} = 1$; there exists $s > 0$ such that $||s\phi||_{H^1}^2 = ||s\phi||_p^p$. We conclude that $s = ||\phi||_p^{\frac{-p}{p-2}}$ and

$$\alpha_M \leq \left(\frac{1}{2} - \frac{1}{p}\right) ||s\phi||_{H^1}^2 = \frac{p-2}{2p} s^2 = \frac{p-2}{2p} ||\phi||_p^{\frac{-2p}{p-2}}.$$

Therefore $||\phi||_p \leq (\frac{2p}{p-2}\alpha_M)^{\frac{2-p}{2p}}$ and

$$\begin{aligned} |f_n(\phi)| &= \left| \int_\Omega |u_n|^{p-2} u_n \phi \right| \leq \left(\int_\Omega |u_n|^p \right)^{\frac{p-1}{p}} \left(\int_\Omega |\phi|^p \right)^{\frac{1}{p}} \\ &\leq \left(\frac{2p}{p-2}\alpha_M \right)^{\frac{p-1}{p}} \left(\frac{2p}{p-2}\alpha_M \right)^{\frac{2-p}{2p}} + o(1) = \left(\frac{2p}{p-2}\alpha_M \right)^{\frac{1}{2}} + o(1) \text{ as } n \to \infty. \end{aligned}$$

We have

$$||f_n||_{H^{-1}} \leq \left(\frac{2p}{p-2}\alpha_M \right)^{\frac{1}{2}} + o(1) \text{ as } n \to \infty. \tag{2.2}$$

Furthermore,

$$f_n\left(\frac{u_n}{\|u_n\|_{H^1}} \right) = \frac{b(u_n)}{a(u_n)^{1/2}} = b(u_n)^{1/2} = \left(\frac{2p}{p-2}\alpha_M \right)^{\frac{1}{2}} + o(1) \text{ as } n \to \infty. \tag{2.3}$$

By (2.2) and (2.3) we conclude that

$$\|f_n\|_{H^{-1}} = \left(\frac{2p}{p-2}\alpha_M\right)^{\frac{1}{2}} + o(1) \text{ as } n \to \infty. \tag{2.4}$$

By the Riesz representation theorem, for each n there is $w_n \in H_0^1(\Omega)$ such that, for each $\varphi \in H_0^1(\Omega)$,

$$f_n(\varphi) = \langle w_n, \varphi\rangle = \int_\Omega (\nabla w_n \cdot \nabla\varphi + w_n\varphi),$$

$$||w_n||_{H^1} = ||f_n||_{H^{-1}} = \left(\frac{2p}{p-2}\alpha_M\right)^{\frac{1}{2}} + o(1) \text{ as } n \to \infty. \tag{2.5}$$

We conclude that

$$\langle w_n, u_n\rangle = f_n(u_n) = \int_\Omega |u_n|^p = \frac{2p}{p-2}\alpha_M + o(1) \text{ as } n \to \infty. \tag{2.6}$$

By (2.1), (2.5), and (2.6) we obtain

$$\begin{aligned} \|u_n - w_n\|_{H^1}^2 &= \|u_n\|_{H^1}^2 - 2\langle u_n, w_n\rangle + \|w_n\|_{H^1}^2 \\ &= \frac{2p}{p-2}\alpha_M - 2\frac{2p}{p-2}\alpha_M + \frac{2p}{p-2}\alpha_M + o(1) \\ &= o(1) \text{ as } n \to \infty. \end{aligned}$$

For $\varphi \in H_0^1(\Omega)$, $||\varphi||_{H^1} = 1$, we have

$$\begin{aligned} \langle J'(u_n), \varphi\rangle &= \int_\Omega (\nabla u_n \cdot \nabla\varphi + u_n\varphi) - \int_\Omega |u_n|^{p-2}u_n\varphi \\ &= \langle u_n, \varphi\rangle - \langle w_n, \varphi\rangle = \langle u_n - w_n, \varphi\rangle, \end{aligned}$$

so

$$|\langle J'(u_n), \varphi\rangle| \le ||u_n - w_n||_{H^1}.$$

We conclude that

$$J'(u_n) = o(1) \text{ strongly in } H^{-1}(\Omega) \text{ as } n \to \infty. \qquad \square$$

In order to study the number β in $\mathbf{F}$, we study the Nehari manifold $\mathbf{M}$ through the unit sphere $\mathbf{S}$ and the zero energy manifold $\mathbf{Z}$ defined by

$$\begin{aligned} &\mathbf{S} = \left\{u \in H_0^1(\Omega) \,|\, \|u\|_{H^1} = 1\right\}, \\ &\mathbf{Z} = \left\{u \in H_0^1(\Omega) \setminus \{0\} \,\middle|\, \tfrac{1}{2}a(u) = \tfrac{1}{p}b(u)\right\}. \end{aligned}$$

Note that $\mathbf{M}$ contains every solution of (1.1). We claim that $\mathbf{M}$ and $\mathbf{Z}$ are $C^{1,1}$ isomorphic to the unit sphere $\mathbf{S}$. In fact, for $\lambda \ge 0$, $u \in \mathbf{S}$, let

$$h_u(\lambda) = J(\lambda u) = \frac{1}{2}\lambda^2 a(u) - \frac{1}{p}\lambda^p b(u).$$

Then

$$\begin{cases} h_u'(\lambda) = \lambda a(u) - \lambda^{p-1} b(u), \\ h_u''(\lambda) = a(u) - (p-1)\lambda^{p-2} b(u). \end{cases}$$

From these properties we can take uniquely r_u, s_u, and $t_u \in \boldsymbol{R}_+$ such that $0 < r_u < s_u < t_u$, $s_u u \in \mathbf{M}$, $t_u u \in \mathbf{Z}$, and

$$0 = h_u''(r_u) = h_u'(s_u) = h_u(t_u).$$

Let $m : \mathbf{S} \to \mathbf{M}$ and $z : \mathbf{S} \to \mathbf{Z}$ be given by $m(u) = s_u u$ and $z(u) = t_u u$. We apply the implicit function theorem and the Sobolev imbedding theorem to obtain the following.

Lemma 2.2.

(1) *m is bijective and of $C^{1,1}$. Moreover $\mathbf{M}$ is path-connected and there exists a constant $c > 0$ such that, for $u \in \mathbf{M}$, $||u||_{H^1} \geq c$ and $J(u) \geq c$;*

(2) *z is bijective and of $C^{1,1}$. Moreover $\mathbf{Z}$ is path-connected and there exists a constant $c' > 0$ such that, for $u \in \mathbf{Z}$, $||u||_{H^1} \geq c'$.*

In the following two lemmas, we assert that every number $\beta > 0$ in $\mathbf{F}$ admits several interesting properties.

Lemma 2.3. *Let $\{u_n\} \subset H_0^1(\Omega)$ be a $(PS)_\beta$-sequence for J with $\beta > 0$. Then there is a sequence $\{s_n\}$ in $\boldsymbol{R}_+$ such that $\{s_n u_n\} \subset \mathbf{M}$ and $J(s_n u_n) = \beta + o(1)$ as $n \to \infty$.*

Proof. By Lemma 2.2(1), there is a sequence $\{s_n\}$ in $\boldsymbol{R}_+$ such that $\{s_n u_n\} \subset \mathbf{M}$ and $h_n'(s_n) = 0$ for each n. Thus $s_n a(u_n) = s_n^{p-1} b(u_n)$ for each n. That $a(u_n) = b(u_n)+o(1)$ as $n \to \infty$ implies $s_n = 1+o(1)$ as $n \to \infty$. Therefore $|J(u_n) - J(s_n u_n)| = o(1)$ as $n \to \infty$, or $J(s_n u_n) = \beta + o(1)$ as $n \to \infty$. □

Lemma 2.4. *Let β be in* $\mathbf{F}$. *Then* (1) *$\beta \geq \alpha_I$;* (2) *$\beta \geq \alpha_M$;* (3) *$\beta \geq \alpha_\Gamma$.*

Proof. Let $\{u_n\} \subset H_0^1(\Omega)$ be a $(PS)_\beta$-sequence for J with $\beta > 0$: that is,

$$\begin{cases} \frac{1}{2} a(u_n) - \frac{1}{p} b(u_n) = \beta + o(1) \text{ as } n \to \infty, \\ a(u_n) - b(u_n) = o(1) \text{ as } n \to \infty. \end{cases}$$

Then $\{u_n\}$ is bounded in $H_0^1(\Omega)$ and $(\frac{1}{2} - \frac{1}{p}) a(u_n) = \beta + o(1)$ as $n \to \infty$.

(1) Let $w_n = u_n (b(u_n))^{-1/p}$, then $b(w_n) = 1$ and $a(w_n) = a(u_n) b(u_n)^{-2/p} \geq I$. Thus $a(u_n) \geq I^{p/(p-2)} + o(1)$ as $n \to \infty$, or $\beta \geq (\frac{1}{2} - \frac{1}{p}) I^{p/(p-2)} = \alpha_I$.

(2) By Lemma 2.3, there is a sequence $\{s_n\}$ in $\boldsymbol{R}_+$ such that $\{s_n u_n\} \subset \mathbf{M}$ and $J(s_n u_n) = \beta + o(1)$ as $n \to \infty$. Therefore $\beta \geq \alpha_M$.

(3) By Lemma 2.3, there is a sequence $\{s_n\}$ in $\boldsymbol{R}_+$ such that $\{s_n u_n\} \subset \mathbf{M}$ and $J(s_n u_n) = \beta + o(1)$ as $n \to \infty$. By Lemma 2.2 (2) there is a sequence $\{t_n\}$ in $\boldsymbol{R}_+$ such that $\{t_n u_n\}$ in $\mathbf{Z}$. Since the manifold $\mathbf{Z}$ is path-connected, there is a path η_n in $\mathbf{Z}$ which connects $t_n u_n$ to e. Let γ_n' be the line segment connecting 0 and $t_n u_n$ and the path $\gamma_n = \gamma_n' \cup \eta_n$. We obtain

$$\alpha_\Gamma \leq \max_{0 \leq t \leq 1} J(\gamma_n(t)) = J(s_n u_n) = \beta + o(1) \text{ as } n \to \infty.$$

Thus $\beta \geq \alpha_\Gamma$. □

By Lemmas 2.1 and 2.4, we have the following theorem.

Theorem 2.5. $\alpha_I = \alpha_M = \alpha_\Gamma$.

Theorem 2.5 has the following two corollaries.

Corollary 2.6. *The minimax number α_Γ is independent of the choice of $e \in \mathbf{Z}$.*

Corollary 2.7. $\alpha_{\Gamma} = \alpha_{\Gamma'}$.

Proof. Since $\Gamma \subset \Gamma'$, we have $\alpha_{\Gamma} \geq \alpha_{\Gamma'}$.

We claim that $K \cap \mathbf{M} \neq \emptyset$ for each $K \in \Gamma'$. If not, let $K \in \Gamma'$ satisfying $K \cap \mathbf{M} = \emptyset$. Let $K_1 = \{u \in K \backslash \{0\} \mid u \in (0, s_u u)\} \cup \{0\}$ and $K_2 = \{u \in K \mid u \in (s_u u, \infty)\}$. Then K_1 and K_2 are nonempty and relatively closed in K, $K_1 \cap K_2 = \emptyset$ and $K_1 \cup K_2 = K$. This contradicts that K is connected.

Assume that $\alpha_{\Gamma} > \alpha_{\Gamma'}$. By Theorem 2.5, $\alpha(\Omega) = \alpha_{\Gamma}$, there exists $K \in \Gamma'$ such that

$$\max_{u \in K} J(u) < \inf_{u \in \mathbf{M}} J(u).$$

Let $u_0 \in K \cap \mathbf{M}$; then we have

$$J(u_0) \leq \max_{u \in K} J(u) < \inf_{u \in \mathbf{M}} J(u) \leq J(u_0),$$

a contradiction. Therefore $\alpha_{\Gamma} = \alpha_{\Gamma'}$. □

We conclude that the numbers α_I, α_M, α_{Γ}, and $\alpha_{\Gamma'}$ are the same and call any one of them the index $\alpha(\Omega)$ of J in Ω. Thus $\alpha(\Omega) \in \mathbf{F}$ and $\mathbf{F} \subset [\alpha(\Omega), \infty)$. In order to understand more about the index $\alpha(\Omega)$, we give the following definition.

Definition 2.8. *Let $\Omega_1 \subset \Omega_2$ be two unbounded domains in $\boldsymbol{R}^N$. We say Ω_2 is a translation-union domain of Ω_1 if there are, for $i = 1, 2, \ldots, l$, $\tau_i \in \boldsymbol{R}^N$ with $\|\tau_i\| = 1$, and sequences $\{r_n^i\}$ of positive numbers, with $r_n^i \to \infty$ as $n \to \infty$ such that*

$$\Omega_2 = \cup_{i=1}^{l} \cup_{n=1}^{\infty} (\Omega_1 + r_n^i \tau_i).$$

Example 2.9.

(1) *Let $\Omega \subset \boldsymbol{R}^N$ be a ball-up domain: that is to say, that for any $r > 0$ there exists $x \in \Omega$ such that $B(x; r) \subset \Omega$. Then $\boldsymbol{R}^N$ is a translation-union domain of Ω.*

(2) *Given $r > 0$, $s \in \boldsymbol{R}$, let*

$$\begin{aligned}
&A^r = \{(x_1, x_2, \ldots, x_N) \in \boldsymbol{R}^N \mid x_1^2 + \cdots + x_{N-1}^2 < r^2\};\\
&A_s^r = \{(x_1, x_2, \ldots, x_N) \in \boldsymbol{R}^N \mid x_1^2 + \cdots + x_{N-1}^2 < r^2, x_N > s\};\\
&A^r \backslash \omega, \text{ where } \omega \subset A^r \text{ is a bounded domain;}\\
&D^r = \{(x_1, x_2, \ldots, x_N) \in \boldsymbol{R}^N \mid x_2^2 + \cdots + x_N^2 < r^2\};\\
&D_s^r = \{(x_1, x_2, \ldots, x_N) \in \boldsymbol{R}^N \mid x_2^2 + \cdots + x_N^2 < r^2, x_1 > s\};\\
&D^r \backslash \omega, \text{ where } \omega \subset D^r \text{ is a bounded domain.}
\end{aligned}$$

Then A^r is a translation-union domain of A_s^r and of $A^r \backslash \omega$; D^r is a translation-union domain of D_s^r and of $D^r \backslash \omega$; and $A^r \cup D^r$ is a translation-union domain of $A_s^r \cup D_s^r$ and of $(A^r \backslash \omega) \cup (D^r \backslash \omega)$.

We have the following important properties.

Proposition 2.10.

(1) *Let $\Omega_1 \subsetneq \Omega_2$ and $J : H_0^1(\Omega_2) \to \boldsymbol{R}$ the energy functional. If J satisfies the $(PS)_{\alpha_1}$-condition or in particular α_1 is a critical value, then $\alpha_2 < \alpha_1$.*

(2) *Let Ω_2 be a translation-union domain of Ω_1. Then $\alpha(\Omega_1) = \alpha(\Omega_2)$, J does not satisfy the $(PS)_{\alpha_1}$-condition, and the only possible solutions of* (1.1) *in Ω_1 are higher energy solutions.*

Proof.

(1) $\Omega_1 \subset \Omega_2$, so $\alpha_2 \leq \alpha_1$. Suppose that J satisfies the $(\mathrm{PS})_{\alpha_1}$-condition; then there exists $u_0 \in \mathbf{M}_1$ such that $u_0 \geq 0$ and $J(u_0) = \alpha_1$. To the contrary, assume $\alpha_2 = \alpha_1$; then $J(u_0) = \alpha_2 = \inf_{u \in \mathbf{M}_2} J(u)$. It is known that every minimizer of the problem $\alpha_2 = \inf_{u \in \mathbf{M}_2} J(u)$ is a critical point of J. Therefore u_0 solves (1.1) in Ω_2. By the maximum principle, $u_0 > 0$ in Ω_2. This contradicts $u_0 \in H_0^1(\Omega_1)$. Therefore $\alpha_2 < \alpha_1$.

(2) $\Omega_1 \subset \Omega_2$, so $\alpha_2 \leq \alpha_1$. Let $\{u_n\} \subset H_0^1(\Omega_2)$ be a minimizing sequence of α_2 :

$$\begin{aligned} J(u_n) &= \alpha_2 + o(1) \text{ as } n \to \infty, \\ a(u_n) &= b(u_n) \text{ for } n = 1, 2, \ldots. \end{aligned}$$

By the definition of translation-union domain, $\Omega_2 = \cup_{i=1}^{l} \cup_{n=1}^{\infty} (\Omega_1 + r_n^i \tau_i)$, it suffices to prove the case $l = 1$. Let $F^n \subset \Omega_1 + r_n \tau$ be a bounded domain, $F^n \nearrow \Omega_2$ as $n \to \infty$ and O_n a bounded and open set in Ω_1 satisfying $F^n - r_n\tau \subset\subset O_n$. Define

$$v_n(x) = \begin{cases} u_n(x + r_n\tau) & \text{if } x \in F^n - r_n\tau, \\ 0 & \text{if } x \notin O_n. \end{cases}$$

Then

$$\begin{aligned} &v_n(x) \in H_0^1(\Omega_1), \\ &a(u_n) = a(v_n) + o(1) \text{ as } n \to \infty, \\ &b(u_n) = b(v_n) + o(1) \text{ as } n \to \infty. \end{aligned}$$

By Lemma 2.2(1), there exists $s_n > 0$ such that $a(s_n v_n) = b(s_n v_n)$, or $s_n = 1 + o(1)$ as $n \to \infty$. $J(s_n v_n) = (\frac{1}{2} - \frac{1}{p}) a(s_n v_n) = (\frac{1}{2} - \frac{1}{p}) s_n^2 a(v_n) = \alpha_2 + o(1)$ as $n \to \infty$. Therefore $\alpha_1 \leq \alpha_2$. We conclude that $\alpha_1 = \alpha_2$. Then we apply the first part to conclude that J does not satisfy the $(\mathrm{PS})_{\alpha_1}$-condition. As in the first part, if u is a ground state solution of (1.1) in Ω_1, then by the maximum principle, u is a positive solution in Ω_2. This contradicts $u \in H_0^1(\Omega_1)$. Therefore the only possible solutions of (1.1) in Ω_1 are higher energy solutions. □

There are some relative properties.

Lemma 2.11.

(1) *There is a ground state solution of* (1.1) *in the infinite strip A^r;*

(2) *Let $\Omega = A^r \backslash \omega$, where $\omega \subset A^r$ is a bounded domain. Then the only possible positive solutions of* (1.1) *in Ω are higher energy solutions;*

(3) *Let $\Omega = A^r \backslash \omega$, where $\omega \subset A^r$ is small and regular. Then there is a positive solution u of* (1.1) *in Ω;*

(4) *There is a $(PS)_\beta$-value β for J such that $\beta > \alpha(\Omega)$.*

Proof.

(1) See Lien–Tzeng–Wang [7, Example 4.3] for the proof.

(2) The proof is by Proposition 2.9(2).

(3) See Hsu–Wang [5, p. 1002] for the proof.

(4) Let u and Ω be as in (3). Set $\beta = J(u)$ and $u_n = u$ for each n to conclude the proof. □

To get further information on the distributions of $\mathbf{F}$, we need the following two important results.

Proposition 2.12. *The only positive solutions of* (1.1) *in $\boldsymbol{R}^N$ are ground state solutions. Moreover the infimum $\alpha(\boldsymbol{R}^N)$ is achieved by a unique positive regular* *ground state solution* $\bar{u} \in H^1(\boldsymbol{R}^N)$ *of* (1.1) *such that* $\bar{u}$ *is spherically symmetric about some point* x_0 *in* $\boldsymbol{R}^N$, $\bar{u}'(r) < 0$ *for* $r = |x - x_0|$, *and*

$$\lim_{r\to\infty} r^{\frac{N-1}{2}} e^r \bar{u}(r) = \gamma > 0,$$
$$\lim_{r\to\infty} r^{\frac{N-1}{2}} e^r \bar{u}'(r) = -\gamma.$$

Proof. See Gidas–Ni–Nirenberg [4] and Kwong [6]. □

LEMMA 2.13 (decomposition lemma). *Let* $\{u_n\} \subset H_0^1(\Omega)$ *be a* $(PS)_\beta$*-sequence for* J. *Then there is a subsequence* $\{u_n\}$, *integer* $l \geq 0$, l *sequences* $\{x_n^i\}_{n=1}^\infty$ *in* $\boldsymbol{R}^N$, $i = 1, \ldots, l$, *function* $\tilde{u}$, *and* w_i *for* $1 \leq i \leq l$ *such that*

$$\begin{aligned}
&-\triangle\tilde{u} + \tilde{u} = |\tilde{u}|^{p-2}\tilde{u} \text{ in } \Omega, \quad \tilde{u} \in H_0^1(\Omega),\\
&-\triangle w_i + w_i = |w_i|^{p-2} w_i \text{ in } \boldsymbol{R}^N, \quad w_i \in H^1(\boldsymbol{R}^N),\\
&u_n = \tilde{u} + \textstyle\sum_{i=1}^l w_i(\cdot - x_n^i) + \mathrm{o}(1) \text{ strongly in } H^1(\boldsymbol{R}^N) \text{ as } n \to \infty,\\
&J(u_n) = J(\tilde{u}) + \textstyle\sum_{i=1}^l J^\infty(w_i) + \mathrm{o}(1) \text{ as } n \to \infty,\\
&|x_n^i| \to \infty, \quad |x_n^i - x_n^j| \to \infty \quad \text{for} \quad 1 \leq i \neq j \leq l, \text{ as } n \to \infty
\end{aligned}$$

where $J^\infty(u) = \frac{1}{2}\int_{\boldsymbol{R}^N}(|\nabla u|^2 + u^2) - \frac{1}{p}\int_{\boldsymbol{R}^N}|u|^p$ *for* $u \in H^1(\boldsymbol{R}^N)$. *In addition, if* $u_n \geq 0$, *then* $\tilde{u} \geq 0$, $w_i > 0$ *for all* $1 \leq i \leq m$, *and each* w_i *can be chosen to be the unique solution* $\overline{u}$ *in Proposition* 2.11.

Proof. See Lions [8], Struwe [10, p. 169], and Lien–Tzeng–Wang [7, Theorem 4.1] for the proof. □

Now we apply Proposition 2.11 and Lemma 2.12 to get the following.

PROPOSITION 2.14. *If* $\beta \in \mathbf{F}$, *then* $\beta = J(u) + m\alpha(\boldsymbol{R}^N)$, *where* u *is a solution of* (1.1) *in* Ω *and* $m = 0, 1, 2, \ldots$.

3. Palais–Smale conditions. In this section, in terms of the index $\alpha(\Omega)$, we give a necessary and sufficient condition for the energy functional J to satisfy the $(\mathrm{PS})_{\alpha(\Omega)}$-condition.

Let $\Omega_0 = \Omega_1 \cup \Omega_2$, where $\Omega_1 \cap \Omega_2$ is bounded, and $\alpha_i = \alpha(\Omega_i)$ be the index of J in Ω_i for $i = 0, 1, 2$. Since $H_0^1(\Omega_i) \subset H_0^1(\Omega_0)$ and $\mathbf{M}_i \subset \mathbf{M}_0$ for $i = 1, 2$, we have $\alpha_0 \leq \min\{\alpha_1, \alpha_2\}$, and the following theorem.

THEOREM 3.1. *The energy functional* J *satisfies the* $(PS)_{\alpha_0}$*-condition if and only if the inequality* $\alpha_0 < \min\{\alpha_1, \alpha_2\}$ *holds. In particular, if the inequality* $\alpha_0 < \min\{\alpha_1, \alpha_2\}$ *holds, then there is a ground state solution* u_0 *of* (1.1) *in* Ω_0.

Proof. The sufficient condition is proven as follows: Suppose that $\alpha_0 < \min\{\alpha_1, \alpha_2\}$. Let $\{u_n\} \subset H_0^1(\Omega_0)$ such that

$$\begin{aligned}
J(u_n) &= \alpha_0 + \mathrm{o}(1) \text{ as } n \to \infty,\\
J'(u_n) &= \mathrm{o}(1) \text{ as } n \to \infty.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{o}(1) &= \langle J'(u_n), u_n\rangle = a(u_n) - b(u_n) \text{ as } n \to \infty,\\
\alpha_0 + \mathrm{o}(1) &= J(u_n) = \left(\frac{1}{2} - \frac{1}{p}\right) a(u_n) + \mathrm{o}(1) \text{ as } n \to \infty.
\end{aligned}$$

Thus we have

$$a(u_n) = b(u_n) + o(1) = \frac{2p}{p-2}\alpha_0 + o(1) \text{ as } n \to \infty.$$

We claim that for each subsequence of $\{u_n\}$ still denoted by $\{u_n\}$, there are $r > 0$, $b > 0$ such that for $Q_r = \Omega_0 \cap B(0, r)$,

$$\int_{Q_r} |u_n|^p \geq b. \tag{3.1}$$

If not, there are $\{r_n\}$, $r_n \to \infty$ and a subsequence $\{u_n\}$ such that for $Q_n = \Omega_0 \cap B(0, r_n)$,

$$\int_{Q_n} |u_n|^p = \mathrm{o}(1) \text{ as } n \to \infty. \tag{3.2}$$

Let $\xi \in C_c^\infty([0, \infty))$ such that

$$0 \leq \xi \leq 1, \quad \xi(t) = \begin{cases} 0 & \text{for } t \in [0, 1], \\ 1 & \text{for } t \in [2, \infty). \end{cases}$$

Let $\xi_n(x) = \xi(\frac{2|x|}{r_n})$ and $w_n = \xi_n^2 u_n$. Since $\{w_n\}$ is bounded in $H_0^1(\Omega_0)$,

$$\begin{aligned} o(1) &= \langle J'(u_n), w_n \rangle \\ &= \int_{\Omega_0} (\xi_n^2 |\nabla u_n|^2 + 2\xi_n u_n \nabla \xi_n \cdot \nabla u_n + \xi_n^2 u_n^2) - \int_{\Omega_0} \xi_n^2 |u_n|^p \text{ as } n \to \infty. \end{aligned}$$

Note that $|\nabla \xi_n(x)| \leq \frac{c}{r_n}$ and (3.2), so

$$\int_{\Omega_0} \xi_n u_n \nabla \xi_n \cdot \nabla u_n = \mathrm{o}(1) \text{ as n} \to \infty,$$

and

$$\int_{\Omega_0} \xi_n^q |u_n|^p = \int_{\Omega_0} |u_n|^p + \mathrm{o}(1) = \frac{2p}{p-2}\alpha_0 + o(1) \text{ as } n \to \infty \text{ for } q > 0. \tag{3.3}$$

We conclude that

$$\int_{\Omega_0} \xi_n^2 (|\nabla u_n|^2 + u_n^2) = \mathrm{o}(1) \text{ as } n \to \infty. \tag{3.4}$$

Let $v_n = \xi_n u_n$. By (3.3) and (3.4),

$$\begin{aligned} J(v_n) &= \frac{1}{2} \int_{\Omega_0} (|\nabla v_n|^2 + v_n^2) - \frac{1}{p} \int_{\Omega_0} |v_n|^p \\ &= \frac{1}{2} \int_{\Omega_0} (|\nabla \xi_n|^2 u_n^2 + \xi_n^2 (|\nabla u_n|^2 + u_n^2) - 2\xi_n u_n \nabla \xi_n \cdot \nabla u_n) \\ &\quad - \frac{1}{p} \int_{\Omega_0} {\xi_n}^p |u_n|^p \\ &= \frac{1}{2} \frac{2p}{p-2}\alpha_0 - \frac{1}{p} \frac{2p}{p-2}\alpha_0 + o(1) \\ &= \alpha_0 + o(1) \text{ as } n \to \infty. \end{aligned}$$

As in the same line of the proof of Lemma 2.1, because $\alpha_0 = \alpha_M$, we have $J'(v_n) = \mathrm{o}(1)$ as $n \to \infty$. Since $\Omega_1 \cap \Omega_2$ is bounded for large n, $v_n = 0$ in $\overline{\Omega_1 \cap \Omega_2}$ and $v_n = v_n^1 + v_n^2$, where $v_n^i \in H_0^1(\Omega_i)$, $i = 1, 2$,

$$v_n^i(x) = \begin{cases} v_n(x) & \text{if } x \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, 2.$$

We obtain

$$J'(v_n^i) = \mathrm{o}(1) \text{ as } n \to \infty \quad \text{for } i = 1, 2.$$

Assume

$$J(v_n^i) = c_i + \mathrm{o}(1) \text{ as } n \to \infty \quad \text{for } i = 1, 2,$$

$$J'(v_n^i) = \mathrm{o}(1) \text{ as } n \to \infty \quad \text{for } i = 1, 2,$$

where $c_1 + c_2 = \alpha_0$. Since c_1 and c_2 are (PS)-values, they are nonnegative. At least one of c_1, c_2 is positive, say $c_1 > 0$. By Lemma 2.4, $c_1 \geq \alpha_1$; thus $\alpha_0 \geq c_1 \geq \alpha_1$. This contradicts $\alpha_0 < \min\{\alpha_1, \alpha_2\}$. Therefore there are $r > 0$, $b > 0$ and for each subsequence $\{u_n\}$ such that for $Q_r = \Omega_0 \cap B(0, r)$,

$$\int_{Q_r} |u_n|^p \geq b.$$

Since $\{u_n\}$ is bounded in $H_0^1(\Omega_0)$, there exists a subsequence $\{u_n\}$ such that

$$u_n \rightharpoonup u_0$$

weakly in $H_0^1(\Omega_0)$, a.e. in Ω_0, and strongly in $L^p_{loc}(\Omega_0)$. Then u_0 is a nonnegative solution of (1.1) in Ω_0. By the fact that $\int_{Q_r} |u_n|^p \geq b$ for each n and by the compact embedding theorem, we have

$$\int_{Q_r} |u_0|^p \geq b.$$

Thus $u_0 \not\equiv 0$. By the maximum principle, u_0 is a positive solution of (1.1) in Ω_0. Thus $u_0 \in M_0$ and

$$J(u_0) \geq \inf_{u \in M_0} J(u) = \alpha_0.$$

Let $p_n = u_n - u_0$; then $\{p_n\}$ is a Palais–Smale sequence for J:

$$\begin{aligned} &J(p_n) = J(u_n) - J(u_0) + \mathrm{o}(1) = \alpha_0 - J(u_0) + \mathrm{o}(1) \text{ as } n \to \infty, \\ &J'(p_n) = \mathrm{o}(1) \text{ as } n \to \infty. \end{aligned}$$

Since $\alpha_0 \geq J(u_0)$, $\alpha_0 = J(u_0)$. Now

$$\mathrm{o}(1) = J(p_n) = \left(\frac{1}{2} - \frac{1}{p}\right) ||p_n||^2_{H^1} + \mathrm{o}(1) \text{ as } n \to \infty.$$

Thus

$$||p_n||^2_{H^1} = \mathrm{o}(1) \text{ as } n \to \infty$$

or

$$u_n \to u_0 \text{ strongly in } H_0^1(\Omega_0) \text{ as } n \to \infty.$$

We conclude that J satisfies the $(\mathrm{PS})_{\alpha_0}$-condition.

The necessary condition is proven as follows: Suppose that $\alpha_0 = \min\{\alpha_1, \alpha_2\}$. Without loss of generality, let $\alpha_0 = \alpha_1$ and $\Omega_1 \subsetneq \Omega_0$. We claim that J does not satisfy the $(\mathrm{PS})_{\alpha_0}$-condition. In fact, suppose on the contrary, J satisfies the $(\mathrm{PS})_{\alpha_0}$-condition in Ω_0. Then we claim that J satisfies the $(\mathrm{PS})_{\alpha_0}$-condition in Ω_1. In fact, let $\{u_n\} \subset H_0^1(\Omega_1)$ satisfy $J(u_n) \to \alpha_1$ and $J'(u_n) \to 0$. There is a subsequence $\{u_n\}$ and $u \in H_0^1(\Omega_0)$ satisfying $u_n \to u$ strongly in $H_0^1(\Omega_0)$; that is to say $u_n \to u$ strongly in $H_0^1(\Omega_1)$. Therefore $J|_{H_0^1(\Omega_1)}$ satisfies the $(\mathrm{PS})_{\alpha_1}$-condition. By Proposition 2.9 (1), $\alpha_0 < \alpha_1$. This is a contradiction. $\square$

4. Solvable and unsolvable domains. In this section, we apply the results in section 3 to solve (1.1) in an unbounded domain Ω.

Esteban–Lions [3, Theorem I.1] proved the following.

Proposition 4.1. *Equation* (1.1) *in an Esteban–Lions domain* Ω *does not admit any nontrivial solution. In particular,* (1.1) *in either* $\boldsymbol{R}_+^N$, *or* A_s^r, *or* D_s^r *does not admit any nontrivial solution.*

We need the following results whose proofs are routine.

Lemma 4.2. *Let* $B_r = \{x \in \boldsymbol{R}^N | \; ||x|| < r\}$, *O be a bounded domain containing* 0 *in* $\boldsymbol{R}^m$, $m \geq 1$, $\Omega = O \times \boldsymbol{R}$, *and* $D_r = r\Omega = \{rx \,|x \in \Omega\}$, $r > 0$. *Then*

(1) $\lim_{r\to\infty} \alpha(B_r) = \alpha(\boldsymbol{R}^N)$;

(2) $\lim_{r\to 0+} \alpha(B_r) = \infty$;

(3) $\lim_{r\to\infty} \alpha(D_r) = \alpha(\boldsymbol{R}^N)$;

(4) $\lim_{r\to 0+} \alpha(D_r) = \infty$.

It is clear that $\alpha(A^r) \leq \alpha(A_s^r)$, $\alpha(D^r) \leq \alpha(D_s^r)$, where $\alpha(A^r)$ and $\alpha(D^r)$ admit minimizers. By Proposition 2.9(2), we obtain the following lemma.

Lemma 4.3. $\alpha(A^r) = \alpha(A_s^r)$ *and* $\alpha(D^r) = \alpha(D_s^r)$ *for any* $s \in \boldsymbol{R}$.

Note that $B_t = \{x \in \boldsymbol{R}^N \mid ||x|| < t\}$ is solvable, but A_0^r is unsolvable since it is an Esteban–Lions domain. However, their union $\Omega_t = B_t \cup A_0^r$, for a fixed r, is solvable.

Theorem 4.4. *There exists* $t_0 > 0$ *such that if* $t \geq t_0$, *then* Ω_t *is solvable.*

Proof. Note that $\alpha(A^r) = \alpha(A_0^r)$. Since $A^r \subsetneq \boldsymbol{R}^N$ is solvable, by Proposition 2.9(1), $\alpha(A_0^r) > \alpha(\boldsymbol{R}^N)$. By Lemma 4.2(1), there exists $t_0 > 0$ such that if $t \geq t_0$, then $\alpha(A_0^r) > \alpha(B_t) > \alpha(\Omega_t)$. Then by Theorem 3.1, Ω_t is solvable. $\square$

Let r, t be the fixed positive numbers, $s \in \boldsymbol{R}$, $x_0 = (0, \ldots, r)$, $\Omega_s = B_t(x_0) \cup D_{-s}^r$, and $D_{-s,s}^r = \{(x_1, x_2, \ldots, x_N) \in \boldsymbol{R}^N \mid x_2^2 + \cdots + x_N^2 < r^2,\ s > x_1 > -s\}$. Then we have the following theorem.

Theorem 4.5. *There exists* $s_0 > r$ *such that if* $s \geq s_0$, *then* Ω_s *is solvable.*

Proof. Since D^r is solvable, we have $\alpha(B_t(x_0) \cup D^r) < \alpha(D^r)$. Similar to Lemma 4.2(1) we obtain $\lim_{s\to\infty} \alpha(B_t(x_0) \cup D_{-s,s}^r) = \alpha(B_t(x_0) \cup D^r)$, so there exists $s_0 > r$ such that if $s \geq s_0$, then $\alpha(B_t(x_0) \cup D_{-s}^r) \leq \alpha(B_t(x_0) \cup D_{-s,s}^r) < \alpha(D^r) = \alpha(D_{-s}^r)$. Since $B_t(x_0)$ is solvable, we have $\alpha(B_t(x_0) \cup D_{-s}^r) < \alpha(B_t(x_0))$. Then by Theorem 3.1, Ω_s is solvable. $\square$

$A_{-\rho}^r$ and D_{-s}^r are Esteban–Lions domains, so they are unsolvable. However, their union is solvable.

Theorem 4.6. *Let* $t < r$ *be fixed. There exists* $s_0 > r$ *such that if* $s \geq s_0$ *and* $\Omega_\rho = A_{-\rho}^r \cup D_{-s}^r$, *then for* $\rho > 0$, Ω_ρ *is solvable.*

Proof. Let s be as in Theorem 4.5, and for $\rho > 0$, $\Omega_\rho = A_{-\rho}^r \cup (B_t(x_0) \cup D_{-s}^r)$. By Theorem 4.5, $B_t(x_0) \cup D_{-s}^r$ is solvable; thus $\alpha(\Omega_\rho) < \alpha(B_t(x_0) \cup D_{-s}^r)$. Note that $\alpha(B_t(x_0) \cup D_{-s}^r) \leq \alpha(D_{-s}^r) = \alpha(A_{-\rho}^r)$, or $\alpha(\Omega_\rho) < \alpha(A_{-\rho}^r)$. Then by Theorem 3.1, for $\rho > 0$, Ω_ρ is solvable. $\square$

REFERENCES

[1] H. Berestycki and P. L. Lions, *Nonlinear scalar field equations*, I. *Existence of ground state*, Arch. Rational Mech. Anal., 82 (1983), pp. 313–345.

[2] H. Brezis and L. Nirenberg, *Remarks on finding critical points*, Comm. Pure Appl. Math., XLIV (1991), pp. 939–963.

[3] M. J. Esteban and P. L. Lions, *Existence and non-existence results for semilinear elliptic problems in unbounded domains*, Proc. Roy. Soc. Edinburgh Sect. A, 93 (1982), pp. 1–12.

[4] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in* $\boldsymbol{R}^n$. Adv. in Math. Supplementary Studies, 7 (1981), pp. 369–402.

[5] T. S. Hsu and H. C. Wang, *A perturbation result of semilinear elliptic equations in exterior strip domains*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 983–1004.

[6] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* $\boldsymbol{R}^n$, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[7] W. C. Lien, S. Y. Tzeng, and H. C. Wang, *Existence of solutions of semilinear elliptic problems on unbounded domains*, Differential Integral Equations, 6 (1993), pp. 1281–1298.

[8] P. L. Lions, *The concentration-compactness principle in the calculus of variations, The locally compact case*, Ann. Inst. H. Poincare Anal. Non Lineare, 1 (1984), Part 1, pp. 109–145, Part 2, pp. 223–283.

[9] P. H. Rabinowitz, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, Regional Conference Series in Mathematics, AMS, Providence, RI, 1986.

[10] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[11] C. A. Stuart, *Bifurcation in* $L^p(\boldsymbol{R}^N)$ *for a semilinear elliptic equation*, Proc. London Math. Soc., 45 (1982), pp. 169–192.

SIAM J. MATH. ANAL.
Vol. 31, No. 1, pp. 166–183

GLOBAL EXISTENCE OF STEADY SUPERSONIC POTENTIAL FLOW PAST A CURVED WEDGE WITH A PIECEWISE SMOOTH BOUNDARY[*]

YONGQIAN ZHANG[†]

Abstract. In this paper we use a modified Glimm scheme to construct a global weak solution to the steady supersonic potential flow past a two-dimensional wedge with a piecewise smooth boundary, small vertex angle, and small total variation of the tangent angle along each side.

Key words. supersonic, wedge, piecewise smooth

AMS subject classifications. 35L65, 76N15

PII. S0036141097331056

1. Introduction.

The problem of steady supersonic flow past a wedge with a smooth boundary has been extensively studied by many authors (for instance, see [1, 2, 3, 6, 9, 11, 14] and references therein). The simple case in which both sides are straight was solved in the book [3] by the shock polar. In [9, 14], applying the theory of quasi-linear hyperbolic systems, Li and Schaeffer give the local existence of steady supersonic flow past the two-dimensional curved wedge with the vertex angle less than the critical value. In [1] Chen extended this result to the case of three-dimensional wedge. Recently Chen established the global existence and asymptotic behavior of steady supersonic flow past a convex combined wedge by making use of hodograph transformation (see [2]).

In this paper we study the potential flow past a two-dimensional wedge with a piecewise smooth boundary. Here, as usual, we consider the case of irrotational and polytropic gas in which the pressure p and the density ρ are related by $p = p(\rho) = A\rho^\gamma$, where A is some positive constant and $\gamma > 1$ is an adiabatic exponent. For simplicity, we study the problem for the half of the wedge, that is, we consider the problem

$$(1.1)\qquad \begin{cases} (\rho u)_x + (\rho v)_y = 0, & \\ v_x - u_y = 0 & \text{in} \quad \Omega \cap \{x > 0\}, \\ (u, v)\cdot \overrightarrow{n} = 0 & \text{on} \quad \partial\Omega, \\ (u, v)|_{x<0} = (q_\infty, 0), & \end{cases}$$

under the following assumptions.

(A1) The Bernoulli relation holds as follows:

$$(1.2)\ \rho = A^{-\frac{1}{\gamma-1}} \left(\frac{\gamma+1}{2\gamma} c_*^2 - \frac{\gamma-1}{2\gamma} q^2 \right)^{\frac{1}{\gamma-1}} = A^{\frac{-1}{\gamma-1}} \left[\frac{\gamma-1}{2\gamma} (q_*^2 - q^2) \right]^{\frac{1}{\gamma-1}}$$

with the constant $q_* = \sqrt{\frac{\gamma+1}{\gamma-1}} c_*$. Here c_* is the critical speed, $q = \sqrt{u^2 + v^2}$, $\gamma > 1$ is the adiabatic exponent, $A > 0$ is some constant.

[*] Received by the editors December 4, 1997; accepted for publication (in revised form) March 5, 1999; published electronically December 3, 1999. This work was partially supported by NSFC Project 19701009 and the Mathematics Center of State Education Commission of China.

http://www.siam.org/journals/sima/31-1/33105.html

[†] Institute of Mathematics, Fudan University, Shanghai 200433, People's Republic of China. Current address: Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan (zhangyq@math.sci.osaka-u.ac.jp).

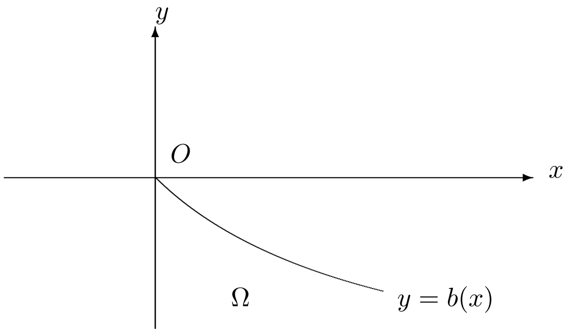


Fig. 1.1.

(A2) The velocity of incoming flow q_∞ is a constant and $q_\infty > c_*$.

(A3) There exists $b \in C[0, +\infty)$ with $b(x) < 0$ for $x > 0$ and $b(0) = 0$ such that $\Omega = \{(x, y)|x \le 0, y < 0\} \cup \{(x, y)|y < b(x), x > 0\}$. In addition there exists a set of points $\{x_k\}_{k=1}^{l} \subseteq (0, +\infty)$ such that $b \in C^\infty[x_{k-1}, x_k]$ for $1 \le k \le l$ and b is affine in $[x_l, +\infty)$. Here $\overrightarrow{n}$ is the outer normal to $\partial\Omega \backslash \{(x_k, b(x_k)), 0 \le k \le l\}$, $x_0 = 0$ (see Figure 1.1).

In [2, 3] the function $y = b(x)$ was assumed to be convex or straight. In this study, we set a rather general assumption on the function $y = b(x)$, that is, there is no special assumption on the shape of the curve $y = b(x)$ except the requirement on the total curvature of the curve. New shock may issue from some place away from the wedge surface; also more complicated boundary interactions may occur (see [3, 5]). To overcome these difficulties we modify the Glimm scheme to handle the initial-boundary value problem. In this paper, under suitable conditions, we shall construct the global weak solution that satisfies $(u, v) = (q_\infty, 0)$ near the line set $\{(x, y)|x = 0, y < 0\}$ and solve the problem (1.1) in the following sense as in [7]:

$$
\begin{aligned}
&\int_{\Omega \cap \{x>0\}} \rho u \phi_{1x} + \rho v \phi_{1y} = \int_{-\infty}^{+\infty} \rho_\infty q_\infty \phi_1(0, y) dy, \\
&\int_{\Omega \cap \{x>0\}} v \phi_{2x} - u \phi_{2y} = 0
\end{aligned}
\tag{1.3}
$$

$\forall \phi_1 \in C_c^\infty(R^2), \phi_2 \in C_c^\infty(\Omega)$ (see [3, 7, 11]). Here $\rho_\infty = \rho(q_\infty, 0)$ is given by (1.2).

The remaining parts of the paper are organized in the following way. In section 2 we first rewrite the equations in the equivalent form and prove that these two systems admit the equivalent entropy conditions. Then by the well-known results of the conservation laws we parameterize the wave curves, i.e., the shock polar and epicycloid in the supersonic region according to the entropy conditions. In section 3 we apply the results in section 2 to establish the existence of solutions to a class of mixed problems and the estimates on the interactions and reflections of waves and the flows past a corner. In section 4 we first approximate the boundary by a collection of straight line segments and modify the Glimm scheme in each approximate domain. In each domain we get the approximate solution and define the Glimm functional analogous to that used in [12, 13] (see also [4, 10]), which is supplemented by additional terms needed to take more complicated boundary interactions into account. The desired decrease of the functional is obtained provided that the top angle and the total curvature of the boundary are sufficiently small. In section 5 we extend the

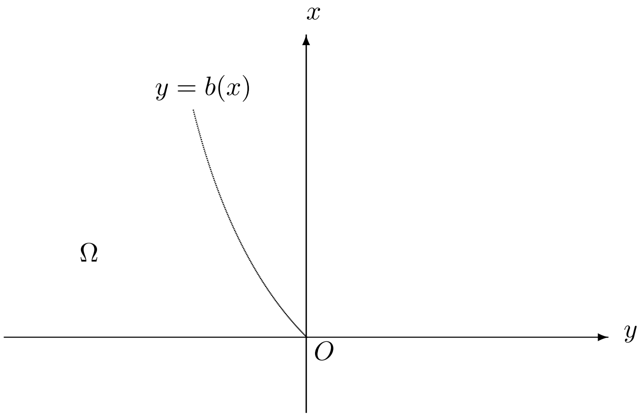


Fig. 2.1.

approximate solutions to the whole domain and prove the compactness of approximate solutions, then obtain the global solution by the convergence of the approximate solutions. Our main results are also stated in section 5.

2. Entropy condition. First we give some notations that will be used throughout the paper. As usual, we view the x-direction as the vertical direction and the y-direction as the horizontal direction and we still use the notation (a, b) to denote the point whose x-coordinate is a and y-coordinate is b (see Figure 2.1).

We recall some basic results about the system. This system is genuinely nonlinear and hyperbolic if the x-direction is regarded as the time direction. It is obvious that the system possesses two distinct characteristics, $\lambda_1 = \frac{uv - c\sqrt{u^2+v^2-c^2}}{u^2-c^2}$, $\lambda_2 = \frac{uv + c\sqrt{u^2+v^2-c^2}}{u^2-c^2}$ and two right eigenvectors $r_j(u, v) = e_j(u, v)\binom{-\lambda_j}{1}$ $(j = 1, 2)$ and $\lambda_1 < 0 < \lambda_2$ near the point $(q_\infty, 0)$. Here $e_j(u, v)$ $(j = 1, 2)$ are smooth functions near the point $(q_\infty, 0)$ which satisfy

$$r_j \cdot \nabla\lambda_j = 1 \tag{2.1}$$

$(j = 1, 2)$ near the point $(q_\infty, 0)$. Moreover we have the following.

Lemma 2.1. *There hold*

$$e_j(u, v) > 0 \tag{2.2}$$

$(j = 1, 2)$ *for any state* (u, v) *near* $(q_\infty, 0)$.

Proof. We prove only the lemma for $j = 2$.

Differentiating the Bernoulli relation $\frac{c^2}{\gamma-1} + \frac{q^2}{2} =$ constant with respect to u and v, we get

$$(c^2)_u|_{(q_\infty,0)} = -(\gamma - 1)q_\infty,$$
$$(c^2)_v|_{(q_\infty,0)} = 0;$$

then

$$\lambda_{2u}|_{(q_\infty,0)} = -\frac{(\gamma-1)q_\infty}{2c_\infty\sqrt{q_\infty^2 - c_\infty^2}} - \frac{(\gamma+1)q_\infty c_\infty}{2\sqrt{(q_\infty^2 - c_\infty^2)^3}} < 0$$

and

$$\lambda_{2v}|_{(q_\infty,0)} = \frac{q_\infty}{q_\infty^2 - c_\infty^2} > 0.$$

Thus it follows

$$\nabla_{(u,v)}\lambda_2 \cdot (-\lambda_2, 1)|_{(q_\infty,0)} > 0.$$

This implies that $e_2 > 0$ near the point $(q_\infty, 0)$. In the same way we can also prove that $e_1 > 0$ near the point $(q_\infty, 0)$. □

Let $R_2(u_0, v_0)$ and $S_2(u_0, v_0)$ (or $R_1(u_0, v_0)$ and $S_1(u_0, v_0)$, resp.) represent, respectively, the epicycloid and shock polar in the supersonic region with respect to λ_2-characteristic field (or λ_1-characteristic field, resp.) passing through (u_0, v_0), $q = \sqrt{u^2 + v^2}$ and denote

$$\begin{aligned} R_2^+(u_0, v_0) &= \{(u, v) \in R_2(u_0, v_0) | q \le q_0.\}, \\ S_2^-(u_0, v_0) &= \{(u, v) \in S_2(u_0, v_0) | q \ge q_0.\}, \\ R_1^+(u_0, v_0) &= \{(u, v) \in R_1(u_0, v_0) | q \ge q_0.\}, \\ S_1^-(u_0, v_0) &= \{(u, v) \in S_1(u_0, v_0) | q \le q_0.\} \end{aligned} \tag{2.3}$$

(see Figure 2.2) and

$$T_j(u_0, v_0) = R_j^+(u_0, v_0) \cup S_j^-(u_0, v_0), \quad j = 1, 2. \tag{2.4}$$

The $T_j(u_0, v_0)$ $(j = 1, 2)$ gives the physically admissible solution with (u_0, v_0) as the left state (see [3, 7]). In addition, from the rotation invariance of the equation and the Hugoniot locus, we have the following lemma.

LEMMA 2.2. *There exists a $\delta_1 > 0$ such that the following holds for all points (u_0, v_0) belonging to the neighborhood of $(q_\infty, 0)$, $O_{\delta_1}((q_\infty, 0))$:*

$$\begin{aligned} R_2^+(u_0, v_0) &= \{(u, v) \in R_2(u_0, v_0) | u \le u_0, v \ge v_0.\}, \\ S_2^-(u_0, v_0) &= \{(u, v) \in S_2(u_0, v_0) | u \ge u_0, v \le v_0.\}, \\ R_1^+(u_0, v_0) &= \{(u, v) \in R_1(u_0, v_0) | u \ge u_0, v \ge v_0.\}, \\ S_1^-(u_0, v_0) &= \{(u, v) \in S_1(u_0, v_0) | u \le u_0, v \le v_0.\}. \end{aligned} \tag{2.5}$$

Set

$$\Psi : \begin{cases} m = \rho u, \\ \quad w = v \end{cases} \tag{2.6}$$

and $W = \binom{m}{w}$, $U = \binom{u}{v}$, $D = \{U \in R^2 | u > c_*, q < q_*\}$.

LEMMA 2.3. *$\Psi : D \longmapsto \Psi(D)$ is a smooth diffeomorphism.*

Proof. A simple calculation shows that

$$m_u = A^{-\frac{1}{\gamma-1}} \left(\frac{\gamma+1}{2\gamma}c_*^2 - \frac{\gamma-1}{2\gamma}q^2\right)^{\frac{2-\gamma}{\gamma-1}} \left(\frac{\gamma+1}{2\gamma}c_*^2 - \frac{\gamma+1}{2\gamma}u^2 - \frac{\gamma-1}{2\gamma}v^2\right) < 0$$

$\forall (u, v) \in D$. This proves the lemma. □

Thus the system can be written in the new coordinates as

$$W_x + H(W)_y = 0, \tag{2.7}$$

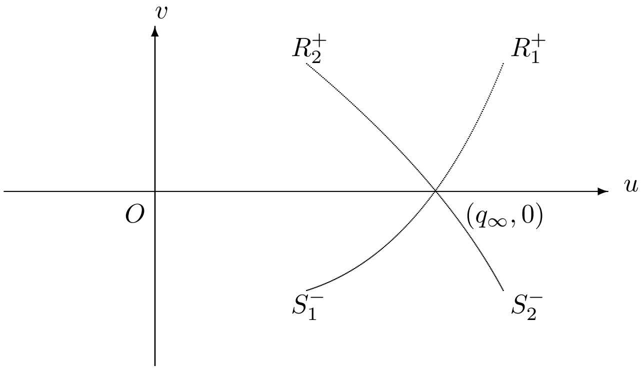


FIG. 2.2. *Wave curves in the case* $u_0 = q_\infty, v_0 = 0$.

where $W = \Psi(u, v)$.

Obviously this new system is also genuinely nonlinear and hyperbolic with respect to the x-direction.

PROPOSITION 2.4. *There is a* $\delta_2 > 0$ *such that the following assertions hold for any state* (u_0, v_0) *near* $U_\infty = (q_\infty, 0)$:

$$\begin{aligned}
&R_j^+(u_0, v_0) \cap O_{\delta_2}(U_\infty) \\
&= \{(u, v) \in R_j(u_0, v_0) | \lambda_j(u, v) \geq \lambda_j(u_0, v_0)\} \cap O_{\delta_2}(U_\infty), \\
&S_j^-(u_0, v_0) \cap O_{\delta_2}(U_\infty) \\
&= \{(u, v) \in S_j(u_0, v_0) | \lambda_j(u, v) \leq \lambda_j(u_0, v_0)\} \cap O_{\delta_2}(U_\infty), \\
&j = 1, 2.
\end{aligned} \tag{2.8}$$

Proof. It suffices to prove the lemma in the case $u_0 = q_\infty$, $v_0 = 0$. First we prove the lemma for S_2^-.

Noticing that S_2 is also the shock curve for the new system, we can parametize S_2 by ϵ with $\frac{dW}{d\epsilon}|_{\epsilon=0} = \tilde{r}_2(W)$ according to Lax [8], where $\tilde{r}_2(W) = \nabla\Psi \cdot r_2|_{U=\Psi^{-1}(W)}$; then the following holds along $S_2(q_\infty, 0)$ by Lemma 2.1:

$$\begin{cases} \dfrac{du}{d\epsilon}\Big|_{\epsilon=0} = -\lambda_2(q_\infty, 0)e_2(q_\infty, 0) < 0, \\ \dfrac{dv}{d\epsilon}\Big|_{\epsilon=0} = e_2(q_\infty, 0) > 0. \end{cases}$$

Therefore it follows that near $\epsilon = 0$

$$\begin{aligned} u(\epsilon) &> u(0) = q_\infty, \\ v(\epsilon) &< v(0) = 0 \end{aligned}$$

holds if and only if $\epsilon < 0$ holds.

Also from (2.1) it follows that near $\epsilon = 0$

$$\lambda_2(u(\epsilon), v(\epsilon)) < \lambda_2(u(0), v(0)) = \lambda_2(q_\infty, 0)$$

holds if and only if $\epsilon < 0$.

This proves the result for $S_2^-(q_\infty, 0)$. The general case for S_2^- follows by the argument of continuity.

The result for S_1^- and $R_{1,2}^+$ can be proved in the same way. □

Equation (2.8) implies that the system in (1.1) and the system (2.7) admit the same entropy condition, that is, they are equivalent in the weak sense. Therefore we can parameterize these waves easily as follows: for any state $W_l \in \Psi(O_{\delta(1)}(U_\infty))$, where $\delta(1) = \min(\delta_1, \delta_2)$, let $L_j(W_l)$ $(j = 1, 2)$ be the curves of Lax (see [8]) parameterized by $\epsilon_j \mapsto \tilde{\Phi}_j(\epsilon_j, W_l)$ with $\tilde{\Phi}_j \in C^2$ and

$$\begin{aligned} &\tilde{\Phi}_j|_{\epsilon_j=0} = W_l, \\ &\frac{\partial \tilde{\Phi}_j}{\partial \epsilon_j}\Big|_{\epsilon_j=0} = \tilde{r}_j(W_l). \end{aligned} \tag{2.9}$$

Here $\tilde{r}_j(W) = \nabla\Psi \cdot r_j|_{U=\Psi^{-1}(W)}$ $(j = 1, 2)$.

By Lemma 2.2 and Proposition 2.4 it's obvious that L_j is constituted by $\Psi(S_j^-)$ and $\Psi(R_j^+)$ $(j = 1, 2)$, and we call the waves given by T_j or $\Psi(T_j)$ the elementary waves or j-wave throughout the paper. Moreover, it follows that $\epsilon_j > 0$ along $\Psi(R_j^+)$ while $\epsilon_j < 0$ along $\Psi(S_j^-)$ $(j = 1, 2)$.

Denote

$$\begin{aligned} &\tilde{\Phi}(\epsilon_2, \epsilon_1, W_l) = \tilde{\Phi}_2(\epsilon_2, \tilde{\Phi}_1(\epsilon_1, W_l)), \\ &\Phi_j = \Psi^{-1} \cdot \tilde{\Phi}_j, \quad j = 1, 2, \end{aligned} \tag{2.10}$$

and

$$\Phi(\epsilon_2, \epsilon_1, (u_l, v_l)) = \Phi_2(\epsilon_2, \Psi \cdot \Phi_1(\epsilon_1, \Psi(u_l, v_l))); \tag{2.11}$$

then we have the following lemma.

Lemma 2.5. *For any pair of states $U_r = \binom{u_r}{v_r}$ and $U_l = \binom{u_l}{v_l}$ close to $U_\infty = \binom{q_\infty}{0}$ in the supersonic region, the system*

$$\begin{cases} W_x + H(W)_y = 0, \\ W|_{x=0} = \begin{cases} W_r, y > 0, \\ W_l, y < 0 \end{cases} \end{cases} \tag{2.12}$$

admits the unique admissible solution constituted by two elementary waves. In addition it owns the representation $(u_r, v_r) = \Phi(\beta, \alpha, (u_l, v_l))$ with

$$\begin{aligned} &\Phi|_{\alpha=\beta=0} = (u_l, v_l), \\ &\frac{\partial \Phi}{\partial \alpha}\Big|_{\alpha=\beta=0} = r_1(u_l, v_l), \end{aligned} \tag{2.13}$$

and

$$\frac{\partial \Phi}{\partial \beta}\Big|_{\alpha=\beta=0} = r_2(u_l, v_l), \tag{2.14}$$

where $W = \binom{\rho u}{v}$.

This lemma can be derived by Lemma 2.2, Proposition 2.4, and the results in the [8]. It will lead to the estimates given in the next section.

For simplicity, we shall use the notation $\{U_l, U_r\} = (\alpha, \beta)$ to denote that $U_r = \Phi(\beta, \alpha, U_l)$ throughout the paper. It is obvious that $\alpha > 0$ along R_1^+ and $\beta > 0$ along R_2^+ while $\alpha < 0$ along S_1^- and $\beta < 0$ along S_2^-.

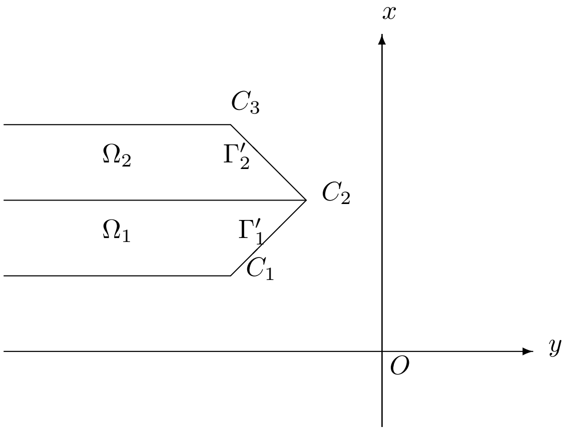


Fig. 3.1.

3. Basic estimates on the nonlinear waves. In this section we shall give the estimates on the interactions and reflections of waves and the flow past the corners. First, by the standard result (see [4, 10]), we have the interaction estimates in the interior as follows.

Lemma 3.1. *If U_l, U_m, and U_r are three states near U_∞, with $\{U_l, U_m\} = \alpha = (\alpha_1, \alpha_2)$, $\{U_m, U_r\} = \beta$, and $\{U_l, U_r\} = \gamma$, then*

$$\gamma_k = \alpha_k + \beta_k + O(1)\Delta'(\alpha, \beta). \tag{3.1}$$

Here $k = 1, 2$ and $\Delta'(\alpha, \beta) = \sum |\alpha_i||\beta_j|$, where the sum is over all pairs for which the ith wave from α and the jth wave from β are approaching; $O(1)$ depends only on the system and U_∞.

Let $C_k = (a_k, b_k)$ $(k = 1, 2, 3)$ with $a_{k+1} > a_k > 0$ $(k = 1, 2)$ (see Figure 3.1) and

$$\begin{aligned}
&\omega = \arctan\frac{b_3 - b_2}{a_3 - a_2} - \arctan\frac{b_2 - b_1}{a_2 - a_1},\\
&\omega_0 = \arctan\frac{b_2 - b_1}{a_2 - a_1},\\
&\Omega_k = \left\{(x, y)|a_k \le x \le a_{k+1}, y < \frac{b_{k+1} - b_k}{a_{k+1} - a_k}(x - a_k) + b_k\right\},\\
&\Gamma'_k = \left\{(x, y)|a_k < x < a_{k+1}, y = \frac{b_{k+1} - b_k}{a_{k+1} - a_k}(x - a_k) + b_k\right\},\\
&\overrightarrow{n_k} = (b_{k+1} - b_k, a_k - a_{k+1}).
\end{aligned}$$

Set

$$\Delta(a, b) = \begin{cases} 0 & \text{if} \quad a \ge 0 \text{ and } \quad b \ge 0,\\ |a||b| & \quad \text{otherwise,}\end{cases}$$

and let H be a neighborhood of U_∞ satisfying $\bar{H} \subset D$ and is compact.

Consider the following mixed problem:

$$\begin{cases} W_x + H(W)_y = 0 & \text{in} \quad \Omega_2,\\ W|_{x=a_2} = W_2, & \\ (u, v) \cdot \vec{n}_2 = 0 & \text{on} \quad \Gamma'_2, \end{cases} \tag{3.2}$$

where $W = \Psi(u, v)$.

LEMMA 3.2. *There exist $\delta_3 > 0$ and $\delta_3' > 0$, $\delta_4 > 0$ with $\delta_4 < \min_H |\arctan \lambda_{1,2}|$ such that if $|U_\infty - U_0| < \delta_3$, $|\omega_0| + |\omega| < \delta_4$ with $U_0 \cdot \overrightarrow{n_1} = 0$, then there exist a unique $\epsilon \in (-\delta_3', \delta_3')$ and a constant state U_2 with $\{U_0, U_2\} = (\epsilon, 0)$ such that the mixed problem (3.2) in Ω_2 with the initial data $W_2 = \Psi(U_0)$ admits a unique admissible solution W constituted by a 1-wave of which the strength is ϵ and $W = \Psi(U_2)$ in some neighborhood of Γ_2'. Moreover,*

$$\epsilon = K_1\omega + O(1)|\omega|^2 \tag{3.3}$$

with $K_1 > c_0 > 0$, where c_0 and the bounds of K_1 and $O(1)$ depend only on the system, U_∞, and $\min_H |\lambda_{1,2}|$.

Proof. It suffices to solve the following equation for the given ω_0, ω, and U_0:

$$\Phi(0, \epsilon, U_0) \cdot (-\sin(\omega_0 + \omega), \cos(\omega_0 + \omega)) = 0. \tag{3.4}$$

Since $\Phi(0, 0, U_\infty) \cdot (0, 1) = 0$ and by Lemma 2.1,

$$\frac{\partial}{\partial \epsilon}[\Phi(0, \epsilon, U_0) \cdot (-\sin(\omega_0 + \omega), \cos(\omega_0 + \omega))] = r_1(q_\infty, 0) \cdot (0, 1) > 0 \tag{3.5}$$

for $\epsilon = \omega = 0$, $\omega_0 = 0$, and $U_0 = U_\infty$, we can get the unique C^2-function $\epsilon = \epsilon(\omega_0 + \omega, U_0)$ which solves the above equation in some neighborhood of $\epsilon = \omega = \omega_0 = 0$ and $U_0 = U_\infty$ by the theorem of implicit function.

Moreover, by assumptions we have

$$\Phi(0, 0, U_0) \cdot (-\sin\omega_0, \cos\omega_0) = 0 \tag{3.6}$$

and this implies $\epsilon(\omega_0, U_0) = 0$. Thus the result follows by the Taylor formula. □

This lemma deals only with the case of the paralleling flow past the corner with small turning angle. To take account of more complicated boundary interaction, including the reflection of waves, we need the following proposition.

PROPOSITION 3.3. *There exist $\delta_i > 0$ $(i = 5, 6)$ and $\delta_5' > 0$ with $\delta_6 < \min_H |\arctan \lambda_{1,2}|$ such that if U_l, U_m, and U_r are three states in the supersonic region with $|U_\infty - U_r| < \delta_5$, $|U_l - U_\infty| < \delta_5$, $|U_m - U_\infty| < \delta_5$, and $|\omega_0| + |\omega| < \delta_6$ and satisfy that $\{U_l, U_m\} = (0, \alpha)$, $\{U_m, U_r\} = (\gamma, 0)$, and $U_r \cdot \overrightarrow{n_1} = 0$, then there exist a unique $\epsilon \in (-\delta_5', \delta_5')$ and a constant state U_2 with $\{U_l, U_2\} = (\epsilon, 0)$ such that the mixed problem (3.2) in Ω_2 with the initial data $W_2 = \Psi(U_l)$ admits a unique admissible solution W constituted by a 1-wave of which the strength is ϵ and $W = \Psi(U_2)$ in some neighborhood of Γ_2'. Moreover,*

$$\epsilon = \gamma + K_3\alpha + K_4\omega + O(1)\{|\alpha||\gamma| + |\alpha||\omega| + \Delta(\gamma, \omega) + |\alpha|^2 + |\omega|^2\} \tag{3.7}$$

holds with $K_3 > 0$ and $K_4 > 0$ and the bounds of K_3, K_4, and $O(1)$ depend only on the system U_∞ and $\min_H |\arctan \lambda_{1,2}|$.

Proof. It suffices to solve the following equations:

$$\begin{aligned} &\Phi(0, \epsilon, U_l) \cdot (-\sin(\omega + \omega_0), \cos(\omega + \omega_0)) \\ &= \Phi(0, \beta, \Phi(0, \gamma, \Phi(\alpha, 0, U_l))) \cdot (-\sin(\omega + \omega_0), \cos(\omega + \omega_0)) \\ &= 0 \end{aligned} \tag{3.8}$$

for the given α, γ, ω, and ω_0 and U_l.

To find the solution (ϵ, β) to the equations, we should carry out the following three steps.

First, noticing $U_r \cdot \vec{n}_1 = 0$, by Lemma 3.2 we can get the unique C^2-function $\beta = \beta(\omega + \omega_0, \Phi(0, \gamma, U_m))$ which solves the equation

$$\Phi(0, \beta, \Phi(0, \gamma, U_m)) \cdot (-\sin(\omega + \omega_0), \cos(\omega + \omega_0)) = 0 \tag{3.9}$$

in the neighborhood of $\beta = \omega_0 = \omega = \gamma = 0$ and $U_m = U_\infty$ and

$$\beta = K_1 \omega + O(1)|\omega|^2, \tag{3.10}$$

with $K_1 > C_0 > 0$.

Also by Lemma 3.1 we can find the unique C^2-function $\epsilon' = \epsilon'(\beta', \gamma, U_m)$ which solves the following equation in some neighborhood of $\epsilon' = \beta = \gamma = 0$ and $U_m = U_\infty$:

$$\Phi(0, \epsilon', U_m) = \Phi(0, \beta', \Phi(0, \gamma, U_m)) \tag{3.11}$$

with

$$\epsilon' = \beta' + \gamma + O(1)\Delta(\beta', \gamma). \tag{3.12}$$

The third step is to solve the following equation:

$$\begin{aligned} &\Phi(0, \epsilon, U_l) \cdot (-\sin(\omega + \omega_0), \cos(\omega + \omega_0)) \\ &= \Phi(0, \epsilon'', \Phi(\alpha, 0, U_l)) \cdot (-\sin(\omega + \omega_0), \cos(\omega + \omega_0)) \end{aligned} \tag{3.13}$$

for the given $\epsilon'', \alpha, \omega, \omega_0$, and U_l.

In the same way as in the proof of Lemma 3.2 we can also get the unique C^2-function $\epsilon = \epsilon(\epsilon'', \alpha, \omega + \omega_0, U_l)$ which solves (3.13) in some neighborhood of $\epsilon = \epsilon'' = \alpha = \omega = \omega_0 = 0$ and $U_l = U_\infty$.

Throughout the paper we omit the U_l in $\epsilon = \epsilon(\epsilon'', \alpha, \omega + \omega_0, U_l)$. It is obvious that

$$\epsilon = I_1 + I_2 + I_3 + I_4 + \epsilon(0, 0, \omega_0), \tag{3.14}$$

where

$$\begin{aligned} I_1 &= \epsilon(\epsilon'', \alpha, \omega + \omega_0) - \epsilon(\epsilon'', 0, \omega + \omega_0) - \epsilon(0, \alpha, \omega + \omega_0) + \epsilon(0, 0, \omega + \omega_0), \\ I_2 &= \epsilon(\epsilon'', 0, \omega + \omega_0), \\ I_3 &= \epsilon(0, \alpha, \omega + \omega_0) - \epsilon(0, 0, \omega + \omega_0) - \epsilon(0, \alpha, \omega_0) + \epsilon(0, 0, \omega_0), \end{aligned}$$

and

$$I_4 = \epsilon(0, \alpha, \omega_0);$$

then we have by the Taylor formula and Lemma 2.1 that

$$\begin{aligned} I_1 &= \mathrm{O}(1)|\epsilon''||\alpha|, \\ I_3 &= \mathrm{O}(1)|\omega||\alpha|, \\ I_4 &= K_3\alpha + \mathrm{O}(1)|\alpha|^2, \end{aligned} \tag{3.15}$$

with $K_3 > 0$. Moreover, it follows by uniqueness that

$$\begin{aligned} &I_2 = \epsilon'', \\ &\epsilon(0, 0, \omega_0) = 0. \end{aligned} \tag{3.16}$$

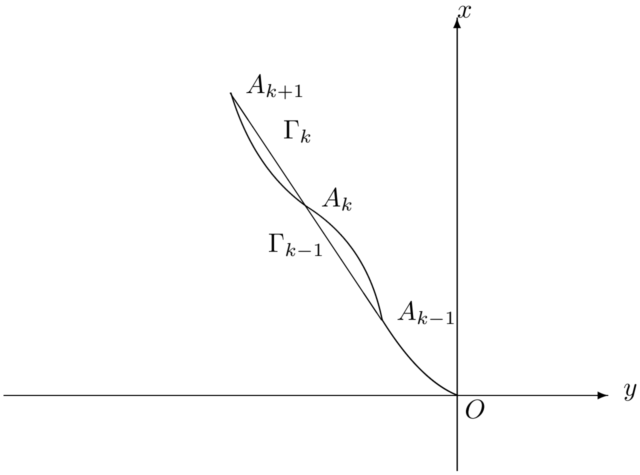


FIG. 4.1.

Hence

$$\epsilon = \epsilon'' + K_3\alpha + O(1)(|\epsilon''||\alpha| + |\omega||\alpha| + |\alpha|^2). \tag{3.17}$$

Let $U_m = \Phi(\alpha, 0, U_l)$, $\epsilon' = \epsilon''$, and $\beta = \beta'$; then by (3.9), (3.11), and (3.13) and noticing

$$\frac{\partial}{\partial\epsilon}\Phi(0, \epsilon, U_l) \cdot (-\sin(\omega + \omega_0), \cos(\omega + \omega_0)) \neq 0 \tag{3.18}$$

for $\epsilon = \omega = \omega_0 = 0$ and $U_l = U_\infty$ we can find the unique ϵ which solves (3.8) in some neighborhood of $\epsilon = \beta = \omega_0 = \omega = \gamma = 0$ and $U_l = U_m = U_r = U_\infty$. In addition, the desired estimates follow from (3.10), (3.12), and (3.17).

The proof is complete. □

4. Glimm scheme. In this section we shall use a modified Glimm scheme to obtain the approximate solution in the approximate domain $\Omega_{\Delta x}$ which will be defined. Without loss of generality we assume that b is smooth and let $y_k = b(k\Delta x)$; we then choose the points $\{A_k = (k\Delta x, y_k)\}_{k=0}^{+\infty}$ in the $\Gamma = \{(x, y)|y = b(x), x \geq 0\}$ and denote

$$\begin{aligned}
&\omega(A_k) = \arctan\frac{y_{k+1} - y_k}{\Delta x} - \arctan\frac{y_k - y_{k-1}}{\Delta x}, \quad k \geq 1,\\
&\omega(A_0) = \arctan\frac{y_1 - y_0}{\Delta x},\\
&\Gamma_k = \{(x, y)|k\Delta x < x < (k + 1)\Delta x, y = b(x, k, \Delta x)\},
\end{aligned}$$

and $\vec{n}_k$ is the outer normal to Γ_k,

$$\Omega_{\Delta x} = \cup_{k\geq 0}\{(x, y)|k\Delta x \leq x < (k + 1)\Delta x, y < b(x, k, \Delta x)\},$$

where $b(x, k, \Delta x) = y_k + \frac{y_{k+1}-y_k}{\Delta x}(x - k\Delta x)$ (see Figure 4.1).

Let $B = \{U \in D||U - U_\infty| < \delta(2)\} \subset H$ and Δy satisfy that $\frac{\Delta y - m\Delta x}{\Delta x} = 2\sup\{|\lambda_{1,2}(z)|, z \in B\}$, where $\delta(2)$ is a constant specialized to meet the requirement of propositions and lemmas in sections 2 and 3, $m = \sup_{k>0}\{\frac{|y_k - y_{k-1}|}{\Delta x}\}$.

Choose mesh points $\{(k\Delta x, a_{k,n})\}_{k\geq 0, -\infty<n<+\infty}$ in R^2 with

$$a_{k,n} = (2n+1+\theta_k)\Delta y + y_k \tag{4.1}$$

and θ_k is randomly chosen in $(-1,1)$. We connect the mesh point $(k\Delta x, a_{k,n})$ by two line segments to the two mesh points, $((k-1)\Delta x, a_{k-1,n-1})$ and $((k-1)\Delta x, a_{k-1,n})$ if $\theta_k \leq 0$, or connect mesh point $(k\Delta x, a_{k,n})$ by two line segments to the two mesh points $((k-1)\Delta x, a_{k-1,n})$ and $((k-1)\Delta x, a_{k-1,n+1})$ if $\theta_k > 0$.

Definition 4.1. *A mesh curve is defined to be an unbounded piecewise linear and space-like curve which is composed of these segments.*

Then each mesh curve I divides the R^2 into I^+ part and I^- part, the I^- part being the one containing $\{x=0\}$. As in [15] we also partially order the mesh curves by saying $I_1 > I_2$ if every point of the mesh curve I_1 is on either I_2 or contained in I_2^+, and call J an immediate successor to I if $J > I$ and every mesh point of J except one is on I.

Now we can define the difference scheme in $\Omega_{\Delta x}$, that is, define the global approximate solution $\tilde{U} = \tilde{U}(x,y)$ in $\Omega_{\Delta x}$. This can be done by carrying out the following steps inductively.

Assume that the approximate solution $\tilde{U}$ has been constructed for $0 \leq x < k\Delta x$ with $\tilde{U}|_{x=0} = U_\infty$ and $\tilde{U}(x,y) = U_j(x,y)$ for $(x,y) \in \{(j-1)\Delta x \leq x < j\Delta x\} \cap \Omega_{\Delta x}$ $(0 \leq j \leq k-1)$, we will define the approximate solution $\tilde{U} = U_k = \Psi^{-1}(W_k)$ in $\{k\Delta x \leq x < (k+1)\Delta x\}$ by solving the following problems.

First we have to solve the following Riemann problem:

$$\begin{cases} (W_k)_x + H(W_k)_y = 0, \\ W_k|_{x=k\Delta x} = W_k^0 \end{cases} \tag{4.2}$$

in each rhombus $T_{k,n}$ whose vertices are $(k\Delta x, (2n-1)\Delta y + y_k)$ $(k\Delta x, (2n+1)\Delta y + y_k)$, $((k+1)\Delta x, (2n-1)\Delta y + y_{k+1})$, $((k+1)\Delta x, (2n+1)\Delta y + y_{k+1})$. Here $n \leq -1$, $W_k^0 = \Psi(U_k^0)$ and

$$U_k^0(y) = U_{k-1}(k\Delta x-, a_{k,n}), \quad y \in (y_k + 2n\Delta y, y_k + 2(n+1)\Delta y).$$

If the problem (4.2) is solvable, define $\tilde{U} = \Psi^{-1}(W_k)$ in $T_{k,n}$ $(n \leq -1)$.

Set

$$U_{k,n} = U_{k-1}(k\Delta x-, a_{k,n}), \quad n \leq -1; \tag{4.3}$$

then by Lemma 3.1 we have that if $W_k^0 \in \Psi(B)$ for $n \leq -1$, the problem (4.2) admits a unique admissible solution and there exist uniquely $\epsilon_{k,n,1}$ and $\epsilon_{k,n,2}$ such that

$$U_{k,n} = \Phi(\epsilon_{k,n,2}, \epsilon_{k,n,1}, U_{k,n-1}). \tag{4.4}$$

Second, to define $\tilde{U}$ in rhombus $T_{k,0}$ whose vertices are $((k+1)\Delta x, y_{k+1})$, $((k+1)\Delta x, \Delta y + y_{k+1})$, $(k\Delta x, \Delta y + y_k)$, and $(k\Delta x, y_k)$, we solve the following mixed problem:

$$\begin{cases} (W_k)_x + H(W_k)_y = 0, \\ W_k|_{x=k\Delta x} = W_k^0, \\ (u_k, v_k)\cdot \overrightarrow{n_k}\,|_{\Gamma_k} = 0 \end{cases} \tag{4.5}$$

in rhombus $T_{k,0}$. If this problem is solvable, then define $\tilde{U} = \Psi^{-1}(W_k)$.

By Lemma 3.2, if $W_k^0 \in \Psi(B)$ and the turning angle is small enough, this problem admits a unique admissible solution; moreover, we can find out a unique $\epsilon_{k,0,1}$ and a constant state $U_{k,0}$ such that

$$U_{k,0} = \Phi(0, \epsilon_{k,0,1}, U_{k,-1}) \tag{4.6}$$

and

$$U_{k,0} \cdot \overrightarrow{n_k}\,|_{\Gamma_k} = 0, \tag{4.7}$$

$$W_k = \Psi(U_{k,0}) \quad \text{in some neighborhood of} \quad \Gamma_k. \tag{4.8}$$

By these solutions we can get the global approximate solution defined in $\Omega_{\Delta x}$. From the discussion above we have the following lemma.

LEMMA 4.2. *For each $k \geq 0$ there exists an $n(k) < 0$ such that*

$$U_{k,n} = U_\infty \quad \forall n < n(k). \tag{4.9}$$

LEMMA 4.3. *If $\{U_l, U_r\} = (\alpha, \beta)$, $U_l, U_r \in B$, then*

$$|U_l - U_r| \leq s(|\alpha| + |\beta|). \tag{4.10}$$

Here $s = \max\{|\frac{\partial}{\partial\alpha}\Phi(\beta,\alpha,U)|, |\frac{\partial}{\partial\beta}\Phi(\beta,\alpha,U)||U \in \bar{H}, |\beta| + |\alpha| \leq \delta_4'\}$.

Throughout the paper we define $U_k(k\Delta x, a_{k,n}) = U_{k,0}$ if $n \geq 0$ for simplification. Then it is obvious that $\epsilon_{k,n,1} = \epsilon_{k,n,2} = 0$ for $n \geq 0$ and $k \geq 0$.

Next we can define the Glimm functional for the approximate solution in $\Omega_{\Delta x}$.

Denote by $U_{\Delta x,\theta}$ the approximate solution constructed above and $W_{\Delta x,\theta} = \Psi(U_{\Delta x,\theta})$, where $\theta = \{\theta_0, \theta_1, \ldots, \theta_k, \ldots\}$. For any mesh curve J, let Ω_J be the set of A_k that lies in J^+, that is,

$$\Omega_J = \{A_k | A_k \in J^+ \cap \partial\Omega_{\Delta x}, A_k = (k\Delta x, y_k)\};$$

denote by α_j (or β_j etc.) the jth wave from α (or β etc.) and by α_j^J (or β_j^J) the strength of α_j (β_j, resp.) wave crossing J ($j = 1, 2$), and denote $K_0 = \sup_B\{K_1, K_2, K_3, K_4\}$ and $K = 8K_0$.

DEFINITION 4.4.

$$\begin{aligned}
L_j(J) &= \sum\{|\alpha_j^J|, \quad \alpha = (\alpha_1, \alpha_2), \alpha_j \textit{ crosses } J\}, \quad j = 1, 2,\\
L_0(J) &= \sum\{|\omega(A)|, \quad A \in \Omega_J\},\\
Q_2(J) &= \sum\{\Delta(\alpha_2^J, \beta_2^J), \quad \alpha_2, \beta_2 \textit{ cross } J \textit{ and } \alpha \textit{ lies to the left of } \beta\},\\
Q_1(J) &= \sum\{\Delta(\alpha_1^J, \beta_1^J), \textit{ both of } \alpha_1, \beta_1 \textit{ cross } J \textit{ and } \alpha \textit{ lies to the left of } \beta\},\\
Q'(J) &= \sum\{|\alpha_2^J||\beta_1^J|, \quad \alpha, \beta \textit{ cross } J, \alpha \textit{ lies to the left of } \beta\},\\
Q''(J) &= \sum\{\Delta(\alpha_1^J, \beta_2^J), \alpha, \beta \textit{ cross } J, \alpha \textit{ lies to the left of } \beta\}\\
&\quad + \sum\{\Delta(\alpha_1^J, \alpha_2^J), \quad \alpha \textit{ crosses } J\},\\
Q_0(J) &= |L_2(J)|^2
\end{aligned}$$

and

$$\begin{aligned} D_2(J) &= \sum\{|\alpha_2^J||\omega(A)|, \quad \alpha_2 \textit{ crosses } J, A \in \Omega_J\} \\ &= L_2(J)L_0(J), \\ D_1(J) &= \sum\{\Delta(\alpha_1^J, \omega(A)), \quad \alpha \textit{ crosses } J, A \in \Omega_J\}, \\ D_0(J) &= \sum\{|\omega(A)|^2, \quad A \in \Omega_J\}, \\ D'(J) &= \sum\{\Delta(\omega(A), \omega(A')), \quad A, A' \in \Omega_J, A \neq A'\}. \end{aligned}$$

Definition 4.5.

$$\begin{aligned} Q(J) &= K^2Q_2(J) + 2KQ'(J) + KQ''(J) + Q_1(J) + K^2Q_0(J), \\ D(J) &= K^2D_2(J) + KD_1(J) + KD_0(J) + K^2D'(J), \\ L(J) &= KL_2(J) + L_1(J) + KL_0(J), \\ F(J) &= L(J) + c\{Q(J) + D(J)\}. \end{aligned}$$

Let $\delta(3) = \min(\delta_4, \delta_6)$ and $\delta(4) = \min(\frac{\delta(2)}{2s}, \frac{\delta_3'}{2}, \frac{\delta_5'}{2})$; then we have the following lemma.

Theorem 4.6. *Let I and J be two mesh curves satisfying $J > I$, and suppose that $|\omega(A_0)| + L_0(x = 0) < \delta(3)$ and I is contained in the domain of definition of $U_{\Delta x,\theta}$ with $U_{\Delta x,\theta}|_I \in B$. There exist constants $c > 0$ and $\delta_7 > 0$ independent of I and J such that if $L(I) < \delta_7$ then J is also contained in the domain of definition of $U_{\Delta x,\theta}$ with $U_{\Delta x,\theta}|_J \in B$ and*

$$F(J) \leq F(I). \tag{4.11}$$

Proof. We first assume J is an immediate successor to I and assume that J and I differ by a single diamond that either lies entirely in the interior of $\Omega_{\Delta x}$ or intersects the boundary of $\Omega_{\Delta x}$.

Case 1. If I and J differ by a single diamond that lies entirely in $\Omega_{\Delta x}$, then $\Omega_I = \Omega_J$. The proof can be carried out in the same way as in [13, 15] by Lemma 3.1. Namely, we can find suitable constants $\delta' \in (0, \delta(4))$ and $c' > 0$ independent of I and J such that if $c \geq c'$ and $L(I) \leq \min(\delta', \frac{1}{(K^2+K+1)c})$, then

$$\begin{aligned} &F(J) \leq F(I), \\ &F(J) \leq L(I) + c(K^2 + K + 1)L(I)^2 \leq 2L(I). \end{aligned}$$

Thus $L(J) \leq \frac{\delta(2)}{s}$ and this implies $U_{\Delta x,\theta}|_J \in B$ by Lemmas 4.2 and 4.3.

Case 2. If I and J differ by a single diamond Λ that intersects the boundary of $\Omega_{\Delta x}$, then Ω_I and Ω_J differ by a single angle ω, that is, $\Omega_I = \Omega_J \cup \{\omega\}$.

Let $I = I_0 \cup I'$ and $J = I_0 \cup J'$ with $J' = \{\epsilon_1\}$; here ϵ_1 is a 1-wave of which the strength crossing J' is ϵ_1. Denote $I_{(2)}$ (or $I_{(1)}$, resp.) the set of 2-waves (or 1-waves, resp.) crossing I. The notations $I_{0,(j)}$, $I'_{(j)}$, $J_{(j)}$, and $J'_{(j)}$ $(j = 1, 2)$ are defined in the same way. In the next notation, without confusion, we shall use $\alpha_j \in I_{*,(j)}$ to denote one j-wave from α of which the strength crossing I_* is α_j.

Define

$$\begin{aligned} Q_1(I_{0,(1)}, \epsilon_1) &= \sum\{\Delta(\beta_1, \epsilon_1), \beta_1 \in I_{0,(1)}\}, \\ Q''(I_{0,(1)}, \alpha_2) &= \sum\{\Delta(\beta_1, \alpha_2), \beta_1 \in I_{0,(1)}\}, \end{aligned}$$

and

$$D_1(I_{0,(1)},\omega)=\sum\{\Delta(\beta_1,\omega),\beta_1\in I_{0,(1)}\}.$$

Now we can carry out the proof. This case is divided into three subcases, for which only the proofs are given.

Subcase (i): If $I'_{(1)}=\{\gamma_1\}$, $I'_{(2)}=\{\alpha_2\}$, α_2 lies to the left of γ_1, then we have

$$\begin{aligned}
&L_2(J)=L_2(I)-|\alpha_2|,\\
&L_0(J)=L_0(I)-|\omega|,\\
&Q_2(J)\le Q_2(I),\\
&Q''(J)=Q''(I_0)=Q''(I)-Q''(I_{0,(1)},\alpha_2),\\
&Q_0(J)=Q_0(I)-|\alpha_2|^2-2|\alpha_2|L_2(I_0),\\
&D_2(J)=D_2(I)-|\alpha_2|L_0(I)-|\omega|L_2(I_0),\\
&D_0(J)=D_0(I)-|\omega|^2,\\
&D'(J)=D'(I)-\sum_{\omega'\in\Omega_J}\Delta(\omega,\omega');
\end{aligned}\tag{4.12}$$

moreover, by Proposition 3.3,

$$\begin{aligned}
L_1(J)\le{}& L_1(I)+K|\alpha_2|+K|\omega|\\
&+\mathrm{O}(1)\{|\alpha_2||\gamma_1|+|\alpha_2||\omega|+\Delta(\gamma_1,\omega)+|\alpha_2|^2+|\omega|^2\},
\end{aligned}\tag{4.13}$$

$$\begin{aligned}
Q_1(J)&=Q_1(I_0)+Q_1(I_{0,(1)},\epsilon_1)\\
&\le Q_1(I)+KQ''(I_{0,(1)},\alpha_2)+KD_1(I_{0,(1)},\omega)\\
&\quad+\mathrm{O}(1)L_1(I_0)\{|\alpha_2||\gamma_1|+|\alpha_2||\omega|+\Delta(\gamma_1,\omega)+|\alpha_2|^2+|\omega|^2\},
\end{aligned}\tag{4.14}$$

$$\begin{aligned}
Q'(J)&=Q'(I_0)+Q'(I_{0,(2)},\epsilon_1)\\
&\le Q'(I)-|\alpha_2||\gamma_1|+K_0|\alpha_2|L_2(I_0)+K_0|\omega|L_2(I_0)\\
&\quad+\mathrm{O}(1)L_2(I_0)\{|\alpha_2||\gamma_1|+|\alpha_2||\omega|+\Delta(\gamma_1,\omega)+|\alpha_2|^2+|\omega|^2\},
\end{aligned}\tag{4.15}$$

and

$$\begin{aligned}
D_1(J)\le{}& D_1(I)+K|\alpha_2|L_0(I)-K|\alpha_2||\omega|-D_1(I_{0,(1)},\omega)\\
&-\Delta(\gamma_1,\omega)+K\sum_{\omega'\in\Omega_J}\Delta(\omega,\omega')\\
&+\mathrm{O}(1)L_0(J)\{|\alpha_2||\gamma_1|+|\alpha_2||\omega|+\Delta(\gamma_1,\omega)+|\alpha_2|^2+|\omega|^2\}.
\end{aligned}\tag{4.16}$$

Thus by (4.12), (4.13), and Definition 4.5 we have the estimate of the linear part,

$$\begin{aligned}
L(J)\le{}& L(I)\\
&+\mathrm{O}(1)K(K+1)\{|\alpha_2||\gamma_1|+|\alpha_2||\omega|+\Delta(\gamma_1,\omega)+|\alpha_2|^2+|\omega|^2\},
\end{aligned}\tag{4.17}$$

and we can get the following estimates of quadratic terms by (4.12), (4.14), (4.15), (4.16), and Definition 4.5:

$$\begin{aligned}
Q(J)\le{}& Q(I)-2K|\gamma_1||\alpha_2|-K^2|\alpha_2|^2\\
&+2K_0K|\omega|L_2(I_0)+KD_1(I_{0,(1)},\omega)\\
&+\mathrm{O}(1)L(I)\{|\alpha_2||\gamma_1|+|\alpha_2||\omega|+\Delta(\gamma_1,\omega)+|\alpha_2|^2+|\omega|^2\}
\end{aligned}\tag{4.18}$$

and

$$\begin{aligned}D(J) \le D(I) &- K^2|\alpha_2||\omega| - K\Delta(\gamma_1,\omega)\\ &- K^2|\omega|L_2(I_0) - K|\omega|^2 - KD_1(I_{0,(1)},\omega)\\ &+ \mathrm{O}(1)L(I)\{|\alpha_2||\gamma_1| + |\alpha_2||\omega| + \Delta(\gamma_1,\omega) + |\alpha_2|^2 + |\omega|^2\};\end{aligned} \tag{4.19}$$

then it follows that

$$\begin{aligned}F(J) \le F(I) &+ (\mathrm{O}(1)L(I)c + \mathrm{O}(1) - (\min(K^2,K))c)\\ &\cdot \{|\alpha_2||\gamma_1| + |\alpha_2||\omega| + \Delta(\gamma_1,\omega) + |\alpha_2|^2 + |\omega|^2\}.\end{aligned} \tag{4.20}$$

Thus we can choose suitable constants $\delta'' \in (0,\delta(4))$ and $c'' > 0$ independent of I and J such that if $c \ge c''$ and $L(I) \le \min(\delta'', \frac{1}{(K^2+K+1)c})$, then

$$\begin{aligned}&F(J) \le F(I),\\ &F(J) \le L(I) + c(K^2+K+1)L(I)^2 \le 2L(I)\end{aligned}$$

and the second inequality implies $U_{\Delta x,\theta}|_J \in B$. This proves subcase (i). There are still two more subcases.

Subcase (ii): If no wave enters Λ, then the same result follows by Lemma 3.2.

Subcase (iii): If there is only a 1-wave entering Λ, the result can be proved in the same way as above by Lemma 3.2 and Proposition 3.3.

Therefore we get the desired result for the case that J is an immediate successor to I. Thus, for the general case, we can pass from I to J by immediate successors, where at each stage F are monotonic nonincreasing and $L \le \min(\delta', \delta'', \frac{1}{(K^2+K+1)c})$ for $c \ge \max(c', c'')$ and where $U_{\Delta x,\theta}$ can be defined and $U_{\Delta x,\theta} \in B$. This proves the desired result. □

Let

$$\omega(0) = \arctan(b'(0)),$$

$$\omega(x_k) = \arctan(b'(x_k+)) - \arctan(b'(x_k-)),$$

and

$$\tau(x) = \frac{b''(x)}{1+(b'(x))^2};$$

then as a corollary of Theorem 4.6 we have the following theorem.

Theorem 4.7. *There exists a* $\delta_0' > 0$ *such that if*

$$|\omega(0)| + \sum_{k=1}^{l}\left(\int_{x_{k-1}}^{x_k} |\tau(x)|dx + |\omega(x_k)|\right) < \delta_0',$$

then there exists a constant $\delta_0'' > 0$ *depending on the function* $b(x)$ *and* δ_0' *such that if* $0 < \Delta x \le \delta_0''$, *then* $U_{\Delta x,\theta}$ *can be defined in* $\Omega_{\Delta x}$ *and Theorem* 4.6 *holds. In addition* $\bigvee_{-\infty}^{y_k}(U_{\Delta x,\theta}(k\Delta x-,\cdot)) \le 3s\delta_0'$ *for any* $k > 0$. *Here the constant* s *is given in Lemma* 4.3 *and* $\bigvee_a^b(w)$ *denotes the total variation of* w *on* $[a,b]$.

5. Convergence of the approximate solution. By (4.8) and (4.9) we can extend $U_{\Delta x,\theta}$ by the constant $U_{k,0}$ continuously across the boundary to the whole strip $\{k\Delta x < x < (k+1)\Delta x\}$ for every $k \geq 0$.

Let the line $\{x = a\}$ intersect $\cup_{k\geq 0}\bar{\Gamma}_k = \cup\{A_{k-1}A_k, k \geq 1\}$ at the point $(a, p_a^{\Delta x})$ for $a > 0$, $y_{n,k} = (2n+1)\Delta x + y_k$, then we have the following lemma.

Lemma 5.1. *The inequality*

$$\int_{-\infty}^{0} |U_{\Delta x,\theta}(x+h, y+p_{x+h}^{\Delta x}) - U_{\Delta x,\theta}(x, y+p_x^{\Delta x})|dy \leq c|h| \tag{5.1}$$

holds for any $h > 0$ and $x > 0$, where the constant c is independent of Δx, θ, and h.

Proof. First by the solution to Riemann problem given in [8], if $k\Delta x \leq x < x + h \leq (k+1)\Delta x$ and $n \leq -1$, we can get

$$\begin{aligned}&\int_{y_{n,k}}^{y_{n+1,k}} |U_{\Delta x,\theta}(x+h, y+p_{x+h}^{\Delta x}) - U_{\Delta x,\theta}(x, y+p_x^{\Delta x})|dy \\ &\qquad \leq c''' \left[\bigvee_{y_{n,k}}^{y_{n+1,k}} (U_{\Delta x,\theta}((k+1)\Delta x-, \cdot))\right] |h|;\end{aligned} \tag{5.2}$$

moreover, by the solution to the mixed problem given in Proposition 3.3,

$$\begin{aligned}&\int_{y_{0,k}}^{y_k} |U_{\Delta x,\theta}(x+h, y+p_{x+h}^{\Delta x}) - U_{\Delta x,\theta}(x, y+p_x^{\Delta x})|dy \\ &\qquad \leq c''' \left[\bigvee_{y_{0,k}}^{y_k} (U_{\Delta x,\theta}((k+1)\Delta x-, \cdot))\right] |h|.\end{aligned} \tag{5.3}$$

Here c''' is a universal constant independent of Δx, θ, and h.

After doing the summation over (5.2) and (5.3), we obtain the estimate by Theorem 4.7.

The general case can be derived by summation over the estimates in each semistrip of $\{k\Delta x \leq x \leq (k+1)\Delta x\}$ and the CFL condition. The proof is complete. □

Lemma 5.2. *If $w \in BV(R^1)$, then*

$$\int_a^b |w(t+h) - w(t)|dt \leq 6 \left[\bigvee_{-\infty}^{+\infty} (w) + |w|_{L^\infty}\right] |h|. \tag{5.4}$$

Proof. It suffices to prove the lemma for $h \geq 0$. Let $g(t) = \bigvee_{-\infty}^{t}(w) - w(t)$ and $f(t) = \bigvee_{-\infty}^{t}(w)$; then $g(t)$ and $f(t)$ are monotonically nondecreasing. Thus we have

$$\begin{aligned}\int_a^b |w(t+h) - w(t)|dt &\leq \int_a^b (f(t+h) + g(t+h))dt - \int_a^b (f(t) + g(t))dt \\ &= \left(-\int_a^{a+h} + \int_b^{b+h}\right)(f(t) + g(t))dt.\end{aligned}$$

This implies (5.4). □

Lemma 5.3. *There holds $|p_{a+h}^{\Delta x} - p_a^{\Delta x}| \leq c|h|$ for any $a \geq 0$ and $h \geq 0$. Here $c = \sup\{|b'(x+)||x \geq 0\}$.*

This result can be derived by direct caculation.

By these lemmas and Lemma 4.2 and Theorem 4.7 we can get the following.

PROPOSITION 5.4. *If $|h|+|l|\le 1$, $D\subseteq\Omega$ is compact, then*

$$\int\int_{D\cap\overline{\Omega}_{\Delta x}}|U_{\Delta x,\theta}(x+h,y+l)-U_{\Delta x,\theta}(x,y)|dxdy\le c(|h|+|l|) \tag{5.5}$$

with the constant c independent of Δx, θ, h, and l.

Set

$$J(\theta,\Delta x,\phi)=\sum_{k=1}^{+\infty}\int_{-\infty}^{0}\phi(k\Delta x,y+y_k)\cdot[U_{\Delta x,\theta}]|_{x=k\Delta x}dy \tag{5.6}$$

with $\phi=(\phi_1,\phi_2)\in C_c^\infty(R^2,R^2)$. Carrying out the same step as in Smoller [15], we have the following proposition.

PROPOSITION 5.5. *There is a null set $N\subset\prod_{k=0}^{+\infty}[-1,1]$ and a sequence $\Delta x_i\xrightarrow[i\longrightarrow+\infty]{}0$, and a $U\in L^1_{loc}(\Omega)\cap L^\infty(\Omega)$ with $U|_{x\le0}=(q_\infty,0)$ such that $J(\theta,\Delta x_i,\phi)\xrightarrow[i\longrightarrow+\infty]{}0$ and $U_{\Delta x_i,\theta}\xrightarrow[i\longrightarrow+\infty]{}U$ strongly in $L^1_{loc}(\Omega\cap\{x\ge0\})$ for any $\theta\in(\prod_{k=0}^{+\infty}[-1,1])\backslash N$ and $\phi_1,\phi_2\in C_c^\infty(R^2)$.*

Now we can establish the global existence.

THEOREM 5.6. *Under the assumptions* (A1), (A2), *and* (A3), *there exists a $\delta_0>0$ such that if $|\omega(0)|+\sum_{k=1}^{l}(\int_{x_{k-1}}^{x_k}|\tau(x)|dx+|\omega(x_k)|)<\delta_0$, the problem* (1.1) *admits a global weak solution in Ω.*

Proof. Choose $\delta_0'>0$ and $\delta_0''>0$ such that Theorem 4.7 holds and let $W=\Psi(U)$ and $U_{\Delta x,\theta}=(u_{\Delta x,\theta},v_{\Delta x,\theta})$ be the approximate solution constructed above.

For any $\phi_2\in C_c^\infty(\Omega)$, there exists a $\delta_0'''\in(0,\delta_0'')$ such that if $\Delta x<\delta_0'''$, then

$$\text{supp}\phi_2\cap\partial\Omega_{\Delta x}=\emptyset. \tag{5.7}$$

So doing the calculation in each rhombus for $\phi_1\in C_c^\infty(R^2)$ and $\phi_2\in C_c^\infty(\Omega)$ and $\Delta x<\delta_0'''$ we have

$$\int_{\Omega_{\Delta x}}W_{\Delta x,\theta}\cdot\phi_x+H(W_{\Delta x,\theta})\phi_y+J(\Delta x,\theta,\phi)=\int_{-\infty}^{0}\rho_\infty q_\infty\phi_1(0,y)dy, \tag{5.8}$$

where $\phi=(\phi_1,\phi_2)$.

Since $|u_{\Delta x,\theta}|\le M$ and $|v_{\Delta x,\theta}|\le M$ for some M by Theorem 4.7 and Lemma 4.2 and

$$\text{mes}(\text{supp}\phi_1\cap\{(\Omega\backslash\Omega_{\Delta x})\cup(\Omega_{\Delta x}\backslash\Omega)\}\cap\{x\ge0\})\xrightarrow[\Delta x\longrightarrow0]{}0,$$

we have

$$\begin{aligned}&\left|\int_{\Omega^+}(W_{\Delta x,\theta}\phi_x+H(W_{\Delta x,\theta})\phi_y)-\int_{\Omega_{\Delta x}}(W_{\Delta x,\theta}\phi_x+H(W_{\Delta x,\theta})\phi_y)\right|\\&\quad\le\int_{(\Omega_{\Delta x}\backslash\Omega^+)\cup(\Omega^+\backslash\Omega_{\Delta x})}|W_{\Delta x,\theta}||\phi_x|+|H(W_{\Delta x,\theta})||\phi_y|\xrightarrow[\Delta x\longrightarrow0]{}0,\end{aligned} \tag{5.9}$$

where $\Omega^+=\Omega\cap\{x\ge0\}$.

Moreover, according to Propositions 5.4 and 5.5, we can find sequences $\Delta x_i \longrightarrow 0$, $\theta \in N$, and U such that $U_{\Delta x_i,\theta} \longrightarrow U$ strongly in L^1_{loc} as $\Delta x_i \longrightarrow 0$ and $J(\theta, \Delta x_i, \phi) \xrightarrow[i \longrightarrow +\infty]{} 0$. Then from (5.8), (5.9), and the discussion in section 4 it follows that U is a weak solution to (1.1) in Ω.

The proof is complete. □

Remark 5.7. In the same way we can also construct a global solution U_+ in Ω_+. Here Ω_+ denote the subdomain of $\{y > 0\}$ that is outside the right half of the wedge. Denote

$$U = \begin{cases} U_+(x, y), & (x, y) \in \Omega_+, \\ U_-(x, y), & (x, y) \in \Omega, \end{cases}$$

and from the structure of the solution we know that U is the desired solution.

Acknowledgment. The author thanks referees for helpful and encouraging comments.

REFERENCES

[1] S. Chen, *Existence of local solution to supersonic flow past a three-dimensional wing*, Adv. Appl. Math., 13 (1992), pp. 273–304.

[2] S. Chen, *Asymptotic behavior of supersonic flow past a convex combined wedge*, Chinese Ann. Math. Ser. B, 19 (1998), pp. 255–264.

[3] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Wiley Interscience, New York, 1948.

[4] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[5] J. Glimm, *The interaction of nonlinear hyperbolic waves*, Comm. Pure Appl. Math., 41 (1988), pp. 569–590.

[6] C. Gu, *A method for solving the supersonic flow past a curved wedge*, Fudan J., 7 (1962), pp. 11–14.

[7] B. Keyfitz and G. Warnecke, *The existence of viscous profiles and admissiblity for transonic shock*, Comm. Partial Differential Equations, 16 (1991), pp. 1197–1221.

[8] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[9] T. Li, *On a free boundary problem*, Chinese Ann. Math., 1 (1980), pp. 351–358.

[10] T. P. Liu, *The deterministic version of the Glimm scheme*, Comm. Math. Phys., 57 (1977), pp. 135–148.

[11] C. S. Morawetz, *On a weak solution for a transonic flow problem*, Comm. Pure Appl. Math., 38 (1985), pp. 797–817.

[12] T. Nishida and J. Smoller, *Mixed problems for nonlinear conservation laws*, J. Differential Equations, 23 (1977), pp. 244–269.

[13] M. Sablè-Tougeron, *Méthod de Glimm et problème mixte*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (4) (1993), pp. 423–443.

[14] D. G. Schaeffer, *Supersonic flow past a nearly straight wedge*, Duke Math. J., 43 (1976), pp. 637–670.

[15] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 31, No. 1, pp. 184–230

WAVELETS ON MANIFOLDS I: CONSTRUCTION AND DOMAIN DECOMPOSITION*

WOLFGANG DAHMEN† AND REINHOLD SCHNEIDER‡

Abstract. The potential of wavelets as a discretization tool for the numerical treatment of operator equations hinges on the validity of norm equivalences for Besov or Sobolev spaces in terms of weighted sequence norms of wavelet expansion coefficients and on certain cancellation properties. These features are crucial for the construction of optimal preconditioners, for matrix compression based on sparse representations of functions and operators as well as for the design and analysis of adaptive solvers. However, for realistic domain geometries the relevant properties of wavelet bases could so far only be realized to a limited extent. This paper is concerned with concepts that aim at expanding the applicability of wavelet schemes in this sense. The central issue is to construct wavelet bases with the desired properties on manifolds which can be represented as the disjoint union of smooth parametric images of the standard cube. The approach considered here is conceptually different though from others working in a similar setting. The present construction of wavelets is closely intertwined with a suitable characterization of function spaces over such a manifold in terms of product spaces, where each factor is a corresponding local function space subject to certain boundary conditions. Wavelet bases for each factor can be obtained as parametric liftings from bases on the standard cube satisfying appropriate boundary conditions. The use of such bases for the discretization of operator equations leads in a natural way to a conceptually new domain decomposition method. It is shown to exhibit the same favorable convergence properties for a wide range of elliptic operator equations covering, in particular, also operators of nonpositive order. In this paper we address all three issues, namely, the characterization of function spaces which is intimately intertwined with the construction of the wavelets, their relevance with regard to matrix compression and preconditioning as well as the domain decomposition aspect.

Key words. topological isomorphisms, Sobolev spaces on manifolds, norm equivalences, complementary boundary conditions, biorthogonal wavelet bases, domain decomposition, boundary integral equations

AMS subject classifications. 46B03, 46E35, 46B15, 65F10, 65N38, 65N55, 65F10, 65F35, 65R20

PII. S0036141098333451

1. Introduction.

1.1. Motivation and perspectives.

Thus far wavelet concepts have unfolded their full computational efficiency mainly when dealing with problems defined on the full Euclidean space or the torus. This is to a great extent due to the fact that in this setting wavelets as discretization tools exhibit some remarkable features.

(I) Wavelet expansions induce *isomorphisms* between *function* and *sequence* spaces [39], that is, certain *Sobolev or Besov norms* of functions are equivalent to weighted *sequence norms* for the coefficients in their wavelet expansions. Specifically, denoting for $s \in \mathbb{R}$ by H^s a scale of Sobolev spaces (possibly incorporating homogeneous boundary conditions), such norm equivalences

*Received by the editors February 2, 1998; accepted for publication (in revised form) February 15, 1999; published electronically December 9, 1999.
http://www.siam.org/journals/sima/31-1/33345.html

†Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, 52056 Aachen, Germany (dahmen@igpm.rwth-aachen.de, http://www.igpm.rwth-aachen.de/∼dahmen/).

‡Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, 09107 Chemnitz, Germany (reinhold.schneider@mathematik.tu-chemnitz.de, http://www.tu-chemnitz.de/∼reinhold/).

have the form

$$c\|\{2^{js}d_{j,k}\}_{j,k}\|_{\ell_2} \leq \|\sum_{j,k} d_{j,k}\psi_{j,k}\|_{H^s} \leq C\|\{2^{js}d_{j,k}\}_{j,k}\|_{\ell_2} \tag{1.1.1}$$

for some range of s.

(II) The wavelets have *cancellation* properties that are usually expressed in terms of *vanishing polynomial moments.*

(I) has immediate important consequences for *preconditioning* systems stemming from elliptic operator equations [19, 17, 32] of positive or even nonnegative order depending on the range of the norm equivalences. In particular, when dealing with operators of negative order, it is important to realize the validity of such norm equivalences as well as for *negative* Sobolev indices $s < 0$ in (1.1.1) in which case the space H^s is understood to be the *dual* of H^{-s}. This latter case has to be treated with some care which we will briefly explain now because this will identify more specific requirements on the wavelet bases.

On the one hand, recall that when $s < 0$, (1.1.1) is proved by establishing an analogous relation for H^{-s} and a *dual* basis $\{\tilde{\psi}_{j,k}\}_{j,k}$. More precisely, suppose that the $\psi_{j,k}$ and $\tilde{\psi}_{j,k}$ are *biorthogonal* with respect to some L_2 inner product $(\cdot,\cdot)$ and that $g := \sum_{j,k} d_{j,k}\psi_{j,k}$ is an element of L_2. Then for (1.1.1) to hold, g has to be identified with a *functional* in H^s. In principle, this can be done through *any* L_2 inner product $\langle\cdot,\cdot\rangle$ by $g(v) := \langle g, v\rangle$ provided that the inner products $(\cdot,\cdot)$ and $\langle\cdot,\cdot\rangle$ are s-equivalent. By this we mean that the Riesz map $R : L_2 \to L_2$ defined by $(\cdot,\cdot) = \langle R\cdot,\cdot\rangle$ not only is an automorphism on L_2 but also extends to one on H^s, i.e., $\|g\|_{H^s}$ and $\|Rg\|_{H^s}$ are equivalent for $g \in L_2$.

On the other hand, the relevance of (1.1.1) for preconditioning stiffness matrices of an operator $\mathcal{L}$ with respect to the wavelet basis hinges on its *H^s-ellipticity.* Specifically, when $\langle\mathcal{L}v, v\rangle$ is equivalent to $\|v\|_{H^s}^2$ for some L_2 inner product $\langle\cdot,\cdot\rangle$, this in turn *determines* how to embed L_2 into H^s for $s < 0$, namely, (up to s-equivalence) through the particular inner product $\langle\cdot,\cdot\rangle$ appearing in the variational formulation of the operator equation [16, 19]. In summary, in order to draw conclusions on preconditioning, it is therefore important to construct biorthogonal wavelet bases not with respect to *any convenient* inner product but to one that is compatible with the underlying variational problem. Since this involves usually the standard L_2 inner product, this is the primary choice considered in this paper.

(II) entails that functions which are smooth except on lower dimensional manifolds have *nearly sparse* wavelet representations. By this we mean that only relatively few coefficients are needed to approximate such a function with desired accuracy. Moreover, applying this principle to the (singular) kernels of a wide class of integral or pseudodifferential operators leads to nearly sparse matrix representations of such operators [5]. This provides the basis for matrix compression schemes whose analysis relies again on (I) and (II). The norm equivalences allow one to transform the continuous problem into a discrete problem that is well posed in the Euclidean metric. In fact, one can show that given the right interplay between the range of norm equivalences and the order of vanishing moments one can, in principle, design efficient solvers which produce approximate solutions with *asymptotically optimal accuracy* at the expense of computational and storage cost that stays *proportional* to the problem size [19, 20, 23, 45].

Again the combination of (I) and (II) (respectively, the consequences with regard to matrix compression) also provides the basis for a rigorous analysis of *adaptive schemes* for elliptic equations. In fact, the analysis of refinement strategies based on *a* *posteriori error estimates* for residuals exploit both (I) and (II) [13, 11]. In particular, convergence in the energy norm can be proved without a priori assumptions on the solution like those commonly needed in a finite element context [6].

Moreover, *nonlinear approximation* is an important theoretical concept related to adaptive approximation. The accuracy that can be achieved by so-called *best N-term approximation* can be characterized in terms of the membership of the approximand to a certain *Besov space* [25]. It is again important to characterize such spaces in terms of discrete norm equivalences.

These facts have motivated various attempts to exploit this potential for the numerical treatment of operator equations. However, the above-mentioned strong implications of wavelet discretizations are valid only under the *assumption* that (I) and (II) hold with appropriate choices of parameters. Unfortunately, as indicated before, so far these properties are conveniently realized within the desired range only when the underlying domain is the full Euclidean space or, via *periodization*, the torus. For more general domain geometries, the construction of appropriate wavelet bases may become prohibitively difficult and expensive.

1.2. Construction principles. Several strategies for dealing with complex domain geometries have been explored in the literature; see [16] for a brief survey and further references. One possible approach is offered by *embedding* techniques. For instance, one can extend the problem to some larger simple domain and enforce the actual boundary conditions by appending them with the aid of Langrange multipliers [36] or correct them by solving a boundary integral equation [3]. However, in both cases a multiresolution setting on the boundary, that is, on a closed manifold, would be highly desirable. This in turn cannot be treated by an embedding strategy.

However, the results in [10, 18] indicate that at least for the *interval*, and hence via tensor products for the unit n-cube, wavelet bases with all the required properties are within reach retaining nearly the full efficiency of wavelet discretizations in the classical setting. It is then fairly straightforward to go one step further. Suppose that $\Omega = \kappa(\square)$, where $\square := (0,1)^n$ and κ is a smooth regular *parametric* mapping. Wavelet bases on $\square$ can then easily be *lifted* to bases on Ω retaining the main driving mechanisms (I) and (II) (see, e.g., [21]). This in turn suggests for us to next consider domains that are *disjoint unions of smooth parametric images* of the standard n-cube $\square$ which will be the setting to be dealt with in this paper.

In fact, in many cases the domain on which the operator equation is defined can be naturally decomposed into a union of simpler domains. For instance, when the domain is a *closed surface*, on which a boundary integral equation is defined, standard CAD packages provide (approximate) representations of such surfaces as a disjoint union of *parametric images* of a standard parameter domain such as the unit square. The individual parametric *patches* are then smoothly joined up to a certain degree of regularity. This means that there exist local reparametrizations for neighboring patches so that the corresponding piecewise defined mapping has a certain number of continuous derivatives; see section 2.1. But this paradigm does not apply only to closed surfaces but also to *bounded domains* (with boundary) in Euclidean space. This is essentially the same point of view as taken in connection with *domain decomposition methods.* Thus a suitable mathematical framework covering all these cases is to view the domain as a (smooth or at least piecewise smooth) *manifold* Γ represented as the union of the disjoint images of some parameter domain. In many cases such as the closed surfaces arising in CAD or domains in CFD the parameter domain can be chosen to be a *cube.*

1.3. Previous approaches and main obstructions. In summary, as pointed out above, the construction of wavelets on manifolds in the above sense has to be intimately connected with the topology of function spaces such as Sobolev and Besov spaces defined on these manifolds. While it is known how to construct suitable bases on each individual patch the problem remains to form from such individual components bases on the *global* manifold which still satisfy (I) and (II). For Sobolev spaces of moderate regularity indices there is no problem. In fact, it is well known that

$$H^s(\Gamma) \asymp \prod_{i=1}^{N} H^s(\Gamma_i), \quad s \in (-1/2, 1/2). \tag{1.3.1}$$

Unfortunately, this is no longer true for $|s| \geq 1/2$. Thus beforehand it is not so clear how to deal with the above task. A natural first idea is to construct a global basis by somehow stitching wavelets defined on the individual patches together so as to realize a certain degree of global smoothness. This idea has been pursued first for special cases in [34, 35] and later in greater generality and in larger range concerning (I) in [21]; see also [7, 12] for slightly different subsequent approaches. However, this concept turns out to have principal limitations. First it requires a *global* parametric representation of the manifold because the wavelets living on more than one parametric patch tie the parametrizations of corresponding adjacent patches together and prohibit local reparametrizations. Hence, aside from expected enormous technical difficulties, a global regularity of a piecewise defined parametrization of higher degree than continuity can only be realized for domains that are topologically equivalent to domains in Euclidean space. Second, in all the above-mentioned cases, pairs of biorthogonal wavelet bases are constructed where biorthogonality is realized with respect to a *modified* L_2-inner product which generally involves *discontinuous* weight functions. Therefore the corresponding Riesz map relating the modified inner product to the standard one (which is simply multiplication by the weight function and hence symmetric) does not take H^s into H^s for $s \geq 1/2$ and, therefore, by duality, neither for $s \leq -1/2$. Hence, on account of the above discussion of (I), whenever ellipticity of the operator is based on using the standard inner product in the variational formulation of the operator equation, relations like (1.1.1) can in this setting be exploited only for preconditioning when $s > -1/2$ which excludes, for instance, the single layer potential operator.

For a restricted class of manifolds including the important case of piecewise affine surfaces with triangular facets, the finite element based wavelets constructed in [24] are indeed biorthogonal bases with respect to the canonical L_2-inner product. This covers the range $|s| \leq 1$, respectively, $|s| \leq 3/2$ for domains in Euclidean space with regard to (I). Moreover, in principle, cancellation properties of any desired order can be realized in this setting, however, at the expense of having explicit local dual bases available.

1.4. Main objectives. The objective of this paper is the construction of biorthogonal wavelet bases with the following properties:

(i) Biorthogonality is realized with respect to some *given* L_2-inner product which in absence of further information will be the canonical one. Both primal and dual wavelets have compact support whose size scales in the usual way.

(ii) The construction applies to manifolds of essentially arbitrary topology.

(iii) Properties (I) and (II) can be realized for *any* range permitted by the regularity of the manifold. In particular, (II) holds in a *patchwise* sense.

Our approach is conceptually different from all the other above mentioned ones. The construction of wavelets will be intimately intertwined with a suitable characterization of function spaces on manifolds. One noteworthy consequence is that a *global* parametric representation of the manifold is never needed so that topology dependent regularity constraints do not arise. We will briefly comment now on these issues.

The basic difficulty is that function spaces on manifolds are usually defined in terms of *open coverings* and associated charts [1], not in terms of *partitions* of the manifold. However, in principle, characterizations of Sobolev and Besov spaces on compact C^∞-manifolds Γ (with or without boundary) of the latter sort have been established in [9]. These results provide the main foundation for the present investigation. The key there is to establish *topological isomorphisms*

$$T : H^s(\Gamma) \to \prod_{i=1}^{N} H^s(\Gamma_i)^\uparrow \tag{1.4.1}$$

between the *global* function space on Γ and a *product* space whose components $H^s(\Gamma_i)^\uparrow$ are corresponding local function spaces defined on the (smooth) patches Γ_i but subject to certain *boundary conditions*. Moreover, in [9] unconditional bases for the individual component spaces were constructed which, with the aid of the previously mentioned isomorphism, lead to discrete norms for the global space. It is important to note that the range of s for which (1.4.1) holds is limited only by the regularity of the manifold.

In full recognition of the fundamental importance of the results in [9] one should note though that the main emphasis has been the *existence* of unconditional bases for function spaces on compact C^∞-manifolds. The existence and structure of the isomorphisms as well as the construction of bases is embedded in a rather involved development. For instance, due to lack of locality and concrete transformation devices which are typical and essential for wavelet schemes, the bases constructed in [9] as well as several constructive ingredients do not yet seem to be practically feasible.

Therefore we will take up the basic concept from [9] here again. Trying first to isolate the relevant ingredients from [9], we realized that, on the one hand, the exposition would be hardly accessible without a complete understanding of [9] and, on the other hand, several crucial deviations from [9] that are necessary from a practical point of view, would not be well founded. Of course, in the above mentioned context one has to deal with less smooth manifolds covering the case of piecewise smooth but globally Lipschitz manifolds.

Thus a first objective of this paper is to rederive topological isomorphisms of the form (1.4.1) in a way that clearly isolates the essential ingredients in a possibly constructive fashion in order to facilitate their adaptation to the computational needs of the concrete problem at hand. A necessary essential prerequisite turns out to be the clear identification of conditions *solely* imposed on certain extension operators so that the rest becomes completely constructive offering clear strategies for further problem dependent modifications. The construction of *scale-dependent* completely *localized* extensions based on suitable local biorthogonal wavelet bases for the parameter domain is one essential distinction of the present approach from the treatment of the desired topological isomorphisms in [9].

The second objective is to reveal the implications of these concepts with regard to the numerical treatment of operator equations. Again appropriate pairs of biorthogonal wavelet bases on the parameter domain play a pivotal role. Together with (1.4.1) they give rise to wavelet bases on the manifold which have optimal localization properties and satisfy requirements (I) and (II) above for any desired range of regularity (permitted by the manifold) and any desired order of cancellation properties. The main consequences for issues like preconditioning and matrix compression will be indicated along with some computational aspects, especially in the context of boundary integral equations. An important point is to reinterpret (1.4.1) as a *domain decomposition method* which appears to differ from those studied in the literature so far and whose convergence properties based on the preceding analysis is now well understood also for operators with global Schwartz kernel.

1.5. Organization of material. Section 2 is devoted to the construction of the isomorphisms T from (1.4.1) which is based on certain projections P_i onto the component spaces. In contrast to [9] we begin with a concrete recursive definition of these projections based on certain *extension* operators from the patches Γ_i to certain neighborhoods. It will be seen that the topological properties of T are completely determined by the topological properties of these extensions and their adjoints.

In section 3 we construct wavelet bases on the manifold which satisfy (I) and (II) for any desired range of regularity permitted by the manifold. We know from [19] that the efficient treatment of boundary integral equations by wavelet schemes requires the option of choosing the order of vanishing moments *higher* than the order of accuracy of the trial spaces. Therefore we employ the concept of *biorthogonal bases* rather than *orthonormal* ones. On account of (1.4.1), the construction of wavelets on the manifold reduces to constructing wavelet bases for the individual component spaces $H^s(\Gamma_i)^{\uparrow}$. Due to the smoothness of the parametric mappings onto each patch Γ_i this can easily be achieved by *lifting* corresponding wavelet bases defined on the unit cube $\square$. At this point we can resort to the results in [22], where exactly those wavelet bases with the right *complementary boundary conditions on the primal and dual side* have been constructed.

Recall that aside from these bases for the component spaces the second ingredient, which the practical feasibility of the approach is ultimately based upon, are suitable extension operators. Therefore special attention will be paid to the realization of appropriate extension operators. Deviating from the developments in [9] we show in section 4 how the multiscale bases on $\square$ can be used to construct *scale-dependent* extension operators that will be seen to significantly improve the efficiency of wavelet schemes for operator equations.

The discretization of operator equations is briefly addressed in section 5. Roughly speaking, (1.4.1) allows one to reformulate a given linear operator equation on Γ as an $N \times N$ *system* of operator equations on the product space. Moreover, when the original operator is self-adjoint and positive definite, the system can be solved iteratively in the spirit of *Schwarz iterations.* In fact, the convergence rate can then be shown to be independent of the discretizations in the individual product spaces provided appropriate wavelet bases are employed. This framework covers differential as well as integral operators. As far as we know this extends the present state of the art for domain decomposition in connection with integral operators significantly. Moreover, due to the validity of (I) and (II), the understanding of adaptive techniques [13, 11] can be fully exploited in this setting. The formulation as a Schwarz iteration has another important practical consequence. For instance, when the operator under consideration only has a global Schwartz kernel, one can choose the extensions in a way that for actual computations the wavelets on Γ *never* have to be determined explicitly. *All* computations refer to problems defined on $\square$ and thus involve wavelet bases defined on $\square$. Moreover, parallel techniques suggest themselves in a natural way.

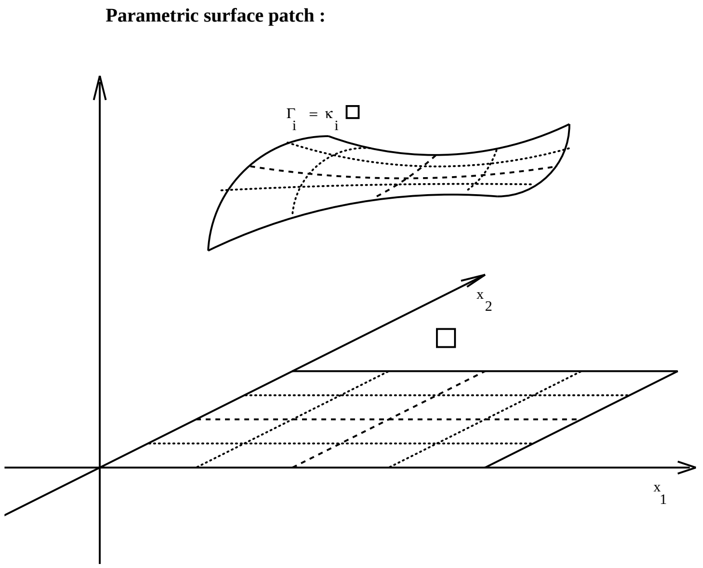


FIG. 1. *Local parametrization.*

2. Function spaces on manifolds.

2.1. Piecewise parametric representations of manifolds. In practical applications, surfaces or manifolds are usually *parametrically* piecewise defined. More precisely, denoting by

$$\square = (0,1)^n,$$

the standard parameter domain, we will assume that

$$\overline{\Gamma}_i = \bigcup_{i=1}^{N} \overline{\Gamma}_i, \quad \Gamma_i = \kappa_i(\square), \quad i = 1, \dots, N, \tag{2.1.1}$$

where

$$\Gamma_i \cap \Gamma_j = \emptyset, \quad i \neq j,$$

and the $\kappa_i : \mathbb{R}^n \to \mathbb{R}^{n'}$, $n \leq n'$ are smooth *regular* parametrizations; see Figure 1. In particular, this means that the induced Lebesgue measures $|\partial\kappa_i(x)|dx = d\mu(\kappa_i(x))$ satisfy

$$c_1 \leq |\partial\kappa_i(x)| \leq c_2, \quad x \in \square \tag{2.1.2}$$

for some positive finite constants c_1, c_2. In most practical situations the regularity of the individual parametric mappings exceeds the global regularity of the manifold. In all currently available surface modeling schemes the κ_i are polynomial or rational mappings of low degree; see, e.g., [31]. The partition of Γ into the patches Γ_i is always supposed to be *conforming*. That is, $\overline{\Gamma}_i \cap \overline{\Gamma}_l$ is either empty or the full parametric image of some lower dimensional face of $\square$ under κ_i and κ_l; see Figure 2.

Moreover we will always assume that Γ is (globally) $C^{m,1}$ for some $m \in \mathbb{N}_0$, $\mathbb{N}_0 := \{0, 1, 2, \dots\}$. Thus Γ coincides locally with the graph of an m times differentiable function with Lipschitz continuous mth order derivatives. It is important to

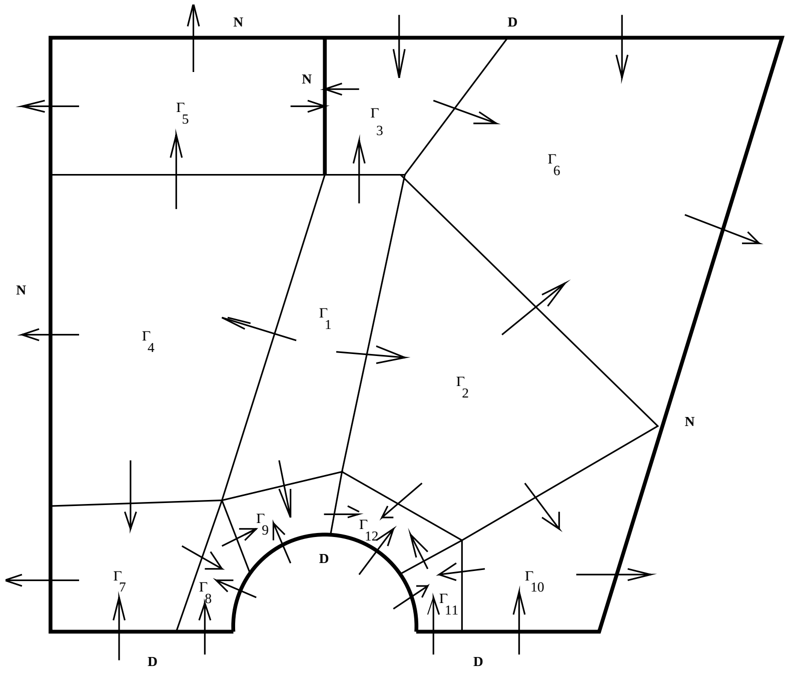


FIG. 2. *Domain with boundary conditions.*

note that in contrast to [12, 7, 21] the individual parmetrizations κ_i are fairly independent of each other. In fact, the factors in the product space on the right-hand side of (1.4.1) are invariant under regular reparametrizations of the patches Γ_i. Therefore each mapping κ_i should rather be viewed as a *representative* of the equivalence class of all parametrizations of Γ_i which are related to κ_i through a $C^{m,1}$-regular reparametrization. The freedom of choosing suitable local reparametrizations will be essential for overcoming topological constraints. Only later in section 4.1 one way of realizing certain extension operators will require a mild local interrelation of the parametrizations of neighboring patches. Roughly speaking, what matters then is the ability of forming for every Γ_i through suitable reparametrizations a piecewise defined $C^{m,1}$-homeomorhic mapping from a neighborhood of $\square$ in $\mathbb{R}^n$ onto a neighborhood of Γ_i; see section 4.1 for more details.

One should also note, however, that the above assumptions are to provide at this point primarily a conceptually convenient basis for the following theoretical developments. In typical practical situations Γ is usually not *given* in the above way but one rather has to *construct* (or approximate) Γ by properly stitching together individual parametric patches in such a way that a certain desired global smoothness is realized. Whenever the patch complex corresponds to a Cartesian grid structure one can employ parametric mappings based on tensor products of univariate sufficiently smooth splines. Of course, in general one encounters *singular vertices*, which means that at such a vertex a number of cubical patches meet that is different from 2^n. It is then much less obvious how to find a piecewise defined parametrization for the union of these patches such that each component of the parametric mapping is $C^{m,1}$. For a general treatment of this question as well as for concrete constructions we refer, for instance, to [30, 27, 31].

The construction of patch complexes such that suitable reparametrizations of some of the patches form local piecewise defined componentwise smooth parametric representations of parts of the surface is a central theme in computer aided geometric design (CAGD) and we will draw upon the techniques developed in this community. This has been primarily developed for surfaces ($n = 2$) which corresponds to the most relevant case with regard to boundary integral equations. Concrete practicable schemes for modeling $C^{m,1}$-surfaces of *arbitrary topology* are by now known for $m = 0, 1, 2$. (Although it is in principle clear how to proceed for higher degrees of smoothness as well.) Typical mild provisions (not imposing any topological constraints) are that singular vertices are sufficiently separated in the patch complex which can always be achieved by dyadic subdivisions of a complex that might initially not meet this requirement; see [31, 42, 44].

2.2. Parametric lifting. We will always assume that Γ is endowed with the induced Lebesgue measure $d\mu$ and $L_0(\Gamma)$ denotes the space of measurable functions on Γ equipped with the topology of convergence in measure on compact sets. By $\langle\cdot,\cdot\rangle_\Gamma$, or more generally, for $\Gamma' \subset \Gamma$, by

$$\langle u, v\rangle_{\Gamma'} = \int\limits_{\Gamma'} u(x)v(x)\,\mathrm{d}\mu(x)$$

we will denote the corresponding L_2-inner product on Γ, Γ', respectively.

We will be concerned with function spaces $\mathcal{F}(\Gamma') \subset L_0(\Gamma')$, $\Gamma' \subset \Gamma$, of the form $\mathcal{F}(\Gamma') = H^s(\Gamma')$ or $\mathcal{F}(\Gamma') = B^s_q(L_p(\Gamma'))$, where $H^s(\Gamma')$, $B^s_q(L_p(\Gamma'))$ denote Sobolev and Besov spaces on Γ', respectively. Here the regularity index s for which these spaces are canonically defined depends on the regularity of the domain Γ'. In particular, under the above assumptions of Γ we have

$$0 \le s \le m + 1 \quad \text{for} \quad m \in \mathbb{N}_0; \tag{2.2.1}$$

see, e.g., [29]. Let s_Γ denote the supremum of all admissible regularity indices. Throughout the following $s < s_\Gamma$ will be fixed in connection with the interpretation of $\mathcal{F}$. The *duals* of $\mathcal{F}(\Gamma')$ (with respect to the duality pairing $\langle\cdot,\cdot\rangle_{\Gamma'}$) will be denoted by $\mathcal{F}^*(\Gamma')$.

The spaces $\mathcal{F}(\Gamma')$ are usually defined via an atlas and partition of unity by lifting corresponding spaces defined on open domains in $\mathbb{R}^n$ [1, 29]; see, e.g., [4, 26] for the definition of these spaces on domains in $\mathbb{R}^n$.

When $\Gamma' = \Gamma_i$, the situation is particularly simple. We record the following observations for later use. To this end, we will briefly write $a \lesssim b$ to express that a can be bounded by a constant multiple of b uniformly with respect to any parameters on which a and b may depend. Moreover, $a \sim b$ means that $a \lesssim b$ and $b \lesssim a$.

Remark 2.2.1. *Let*

$$g_i := |\partial\kappa_i(\kappa_i^{-1}(\cdot))|. \tag{2.2.2}$$

Then setting

$$(u, v)_i := \langle u \circ \kappa_i, v \circ \kappa_i\rangle_\Box, \tag{2.2.3}$$

one has

$$(g_i u, v)_i = \langle u, v\rangle_{\Gamma_i} = \langle|\partial\kappa_i| u \circ \kappa_i, v \circ \kappa_i\rangle_\Box. \tag{2.2.4}$$

By (2.1.2), one has

$$c_1 \leq g_i(x) \leq c_2, \quad x \in \Gamma_i, \tag{2.2.5}$$

so that

$$(v,v)_i \sim \langle v,v\rangle_{\Gamma_i}, \quad v \in L_2(\Gamma_i). \tag{2.2.6}$$

REMARK 2.2.2. *The bilinear form*

$$(u,v) := \sum_{i=1}^{N} (u,v)_i \tag{2.2.7}$$

defines a scalar product for $L_2(\Gamma)$ *such that*

$$\|\cdot\|_{L_2(\Gamma)} \sim (\cdot,\cdot)^{1/2}. \tag{2.2.8}$$

Hence any Riesz basis in $L_2(\Gamma)$ *has a dual with respect to* $(\cdot,\cdot)$ *which also belongs to* $L_2(\Gamma)$.

For any $\Gamma' \subset \Gamma$, the space $\mathcal{F}(\Gamma')$ is defined as the *quotient space* normed by

$$\|g\|_{\mathcal{F}(\Gamma')} := \inf_{f \in \mathcal{F}(\Gamma), f|_{\Gamma'} = g} \|f\|_{\mathcal{F}(\Gamma)}. \tag{2.2.9}$$

We will make use of the following familiar fact; see, e.g., [29].

REMARK 2.2.3. *Assume that* w *is any smooth function on* Γ' *satisfying* (2.2.5). *Then*

$$\|f\|_{\mathcal{F}} \sim \|wf\|_{\mathcal{F}}, \quad f \in \mathcal{F}$$

for any $\mathcal{F}$ *of the form* $\mathcal{F}(\Gamma')$, $\Gamma' \subseteq \Gamma$.

REMARK 2.2.4. *Suppose that* $\mathcal{U}$ *denotes any closed subspace of* $\mathcal{F}$. *Then for any regular parametrization* κ_i *of* Γ_i *one has*

$$\mathcal{U}(\Box) = \{v \circ \kappa_i : v \in \mathcal{U}(\Gamma_i)\} \tag{2.2.10}$$

and

$$\|v\|_{\mathcal{U}(\Gamma_i)} \sim \|v \circ \kappa_i\|_{\mathcal{U}(\Box)}, \quad v \in \mathcal{U}(\Gamma_i). \tag{2.2.11}$$

Of course, regular reparametrizations in (2.2.10) *affect only the constants in* (2.2.11) *and thus give rise to equivalent norms.*

As mentioned before, our objective is to construct topological isomorphisms of the form

$$T : \mathcal{F}(\Gamma) \to \prod_{i=1}^{N} \mathcal{F}(\Gamma_i)^{\uparrow}, \tag{2.2.12}$$

where $\mathcal{F}(\Gamma_i)^{\uparrow}$ are certain closed subspaces of $\mathcal{F}(\Gamma_i)$ which, according to Remark 2.2.4, can be fully described by subspaces of $\mathcal{F}(\Box)$. The superscript $\uparrow$ will be seen to indicate certain boundary conditions as detailed in the next section.

2.3. Numbering of patches and orientation of faces. The construction of T from (2.2.12) involves an appropriate numbering of the patches Γ_i. To construct this numbering it is useful to view the patches as vertices of a graph $\mathcal{G}$. The set of edges $\mathcal{E}$ is identified with the $(n-1)$ faces shared by adjacent patches Γ', Γ''. Any two patches having an $(n-1)$ face in common are called neighbors.

We will construct now a numbering of the vertices of $\mathcal{G}$, and based on this, an *orientation* for $\mathcal{E}$. The numbering for $\mathcal{G}$ will be defined recursively as follows. Pick any patch in $\mathcal{G}$, call it Γ_1, and define

$$\mathcal{G}_1 = \{\Gamma_1\}. \tag{2.3.1}$$

Next form a layer of *level* 2 around $\mathcal{G}_1$ by setting

$$\mathcal{G}_2 = \left\{\Gamma' \in \mathcal{G} \setminus \mathcal{G}_1 \,:\, \overline{\Gamma}_1 \cap \overline{\Gamma}' \in \mathcal{E}\right\}. \tag{2.3.2}$$

We will fix some ordering of $\mathcal{G}_2$ by setting

$$\mathcal{G}_2 = \{\Gamma_{1,i} \,:\, i = 1, \dots, \#\mathcal{G}_2\}. \tag{2.3.3}$$

Suppose now we have constructed subsets $\mathcal{G}_1, \dots, \mathcal{G}_{\ell-1}$ for some $\ell \geq 2$, where $\mathcal{G}_j$ contains all neighbors of the elements in $\mathcal{G}_{j-1} \setminus \mathcal{G}_{j-2}$. Then set

$$\mathcal{G}_\ell = \left\{\Gamma' \in \mathcal{G} \setminus \{\mathcal{G}_1 \cup \dots \cup \mathcal{G}_{\ell-1}\} \,:\, \overline{\Gamma}' \cap \overline{\Gamma}'' \in \mathcal{E},\, \Gamma'' \in \mathcal{G}_{\ell-1}\right\}. \tag{2.3.4}$$

Again assuming that the elements of $\mathcal{G}_{\ell-1}$ are indexed as $\Gamma_{\mathbf{a}}$ by some multi-integer $\mathbf{a} \in \mathbb{N}^{\ell-1}$ we set

$$\mathcal{G}_\ell = \{\Gamma_{\mathbf{a},i} \,:\, \Gamma_{\mathbf{a}} \in \mathcal{G}_{\ell-1},\, \Gamma_{\mathbf{a},i} \text{ a neighbor of } \Gamma_{\mathbf{a}}\}. \tag{2.3.5}$$

Obviously, there exists $L \in \mathbb{N}$ such that $\mathcal{G} = \mathcal{G}_1 \cup \dots \cup \mathcal{G}_L$. Now suppose that we have fixed for each ℓ a total ordering for the elements $\Gamma_{\mathbf{a}}$, $\mathbf{a} \in \mathbb{N}^\ell$, in $\mathcal{G}_\ell$ denoted by $\prec$. For any $\mathbf{a}$, we fix an ordering for the neighbors $\Gamma_{\mathbf{a},i}$, $i = 1, \dots, n_{\mathbf{a}}$ and extend $\prec$ in a lexicographical fashion by

$$(\mathbf{a}, i) \prec (\mathbf{a}', i') \quad \text{iff} \quad \begin{cases} \mathbf{a} \prec \mathbf{a}', & i, i' \text{ arbitrary}, \\ \mathbf{a} = \mathbf{a}', & i < i'. \end{cases} \tag{2.3.6}$$

Obviously this establishes a total ordering in $\mathcal{G}_{\ell+1}$ which immediately extends to a total ordering in $\mathcal{G}$ by

$$\mathbf{a} < \mathbf{a}' \text{ iff } \begin{cases} \ell(\mathbf{a}) < \ell(\mathbf{a}') \text{ or} \\ \ell(\mathbf{a}) = \ell(\mathbf{a}') \text{ and } \mathbf{a} \prec \mathbf{a}', \end{cases} \tag{2.3.7}$$

where $\ell(\mathbf{a})$ is the level ℓ so that $\Gamma_{\mathbf{a}} \in \mathcal{G}_\ell$.

In the following the numbering $(\Gamma_i)_{i=1}^N$ will always be assumed to stem from the above ordering, i.e.,

$$i < j \;\text{ iff }\; \mathbf{a}(i) < \mathbf{a}'(j) \tag{2.3.8}$$

in the sense of (2.3.7).

Each edge $\overline{\Gamma}_i \cap \overline{\Gamma}_l$ in $\mathcal{E}$ will be indexed as $e_{i,l}$ iff $i < l$ which induces an orientation in $\mathcal{E}$. The oriented set of edges will be denoted by $\mathcal{E}^\uparrow$. One may picture this by

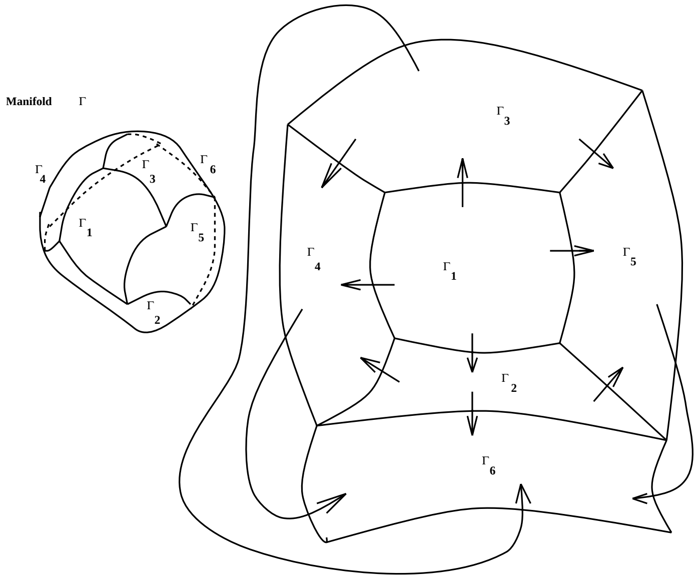


FIG. 3. *Manifold and orientation of patch boundaries.*

associating with $e_{i,l} \in \mathcal{E}^{\uparrow}$ an arrow pointing from the patch Γ_i into the patch Γ_l across the common face; see Figure 3 indicating the decomposition of a closed spherelike surface. The corresponding oriented graph will be denoted by $\mathcal{G}^{\uparrow}$.

The purpose of the above construction is to divide the $(n-1)$ faces of the patches Γ_i into at most two groups, namely, *inflow* or *outflow* faces depending on the orientation. Accordingly, we denote by $\partial^{\uparrow}\Gamma_i$ the *outflow boundary* of the patch Γ_i, i.e.,

$$\partial^{\uparrow}\Gamma_i = \bigcup_l \left\{ e_{i,l} \in \mathcal{E}^{\uparrow} \right\}, \tag{2.3.9}$$

as indicated in Figure 4.

When Γ has a boundary there exist some $(n-1)$ faces which are not yet included in $\mathcal{E}$. We will assign arrows to these boundary faces depending on the type of *boundary conditions* that may be imposed there. If a patch boundary is part of the boundary of Γ, where *homogeneous Dirichlet conditions* are imposed, this edge becomes an inflow boundary while *Neumann boundary conditions* correspond to outflow boundaries; see Figure 2, where respective boundary segments are flagged with D and N for Dirichlet and Neumann conditions, respectively. The rationale behind this will become clear from the subsequent discussion.

We will have to consider *extensions* across the outflow boundary. Accordingly, we need to define an appropriate set of *outflow neighbors*

$$\mathcal{N}_i^{\uparrow} := \left\{ \Gamma_j \in \mathcal{G}^{\uparrow} \,:\, j > i,\, \overline{\Gamma}_j \cap (\text{rel int } \partial^{\uparrow}\Gamma_i) \neq \emptyset \right\}, \tag{2.3.10}$$

which consists of those patches whose boundary intersects the relative interior of the outflow boundary of Γ_i. In complete analogy we define the *inflow boundary*

$$\partial^{\downarrow}\Gamma_i = \bigcup \left\{ e_{\ell,i} \in \mathcal{E}^{\uparrow} \right\} \tag{2.3.11}$$

and its set of neighbors $\mathcal{N}_i^{\downarrow}$

$$\mathcal{N}_i^{\downarrow} = \left\{ \Gamma_\ell \in \mathcal{G}^{\uparrow} \,:\, \ell < i,\, \overline{\Gamma}_j \cap (\text{rel int } \partial^{\downarrow}\Gamma_i) \neq \emptyset \right\}. \tag{2.3.12}$$

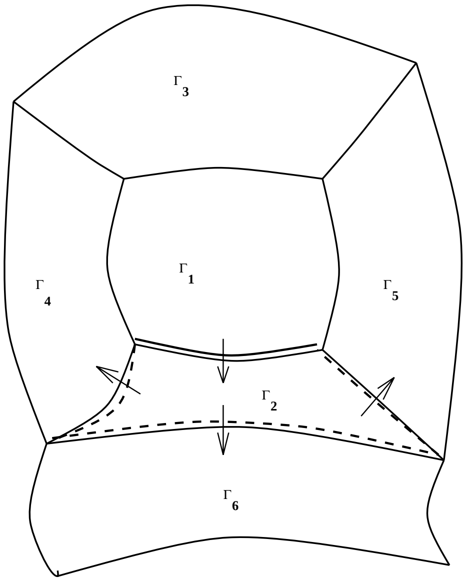


FIG. 4. *Inflow and outflow boundaries.*

With each Γ_i we associate next the open set $\Gamma_i^{\uparrow}$ defined as

$$\Gamma_i^{\uparrow} := \operatorname{int}\left(\overline{\Gamma}_i \bigcup \left\{\overline{\Gamma}' : \Gamma' \in \mathcal{N}_i^{\uparrow}\right\}\right). \tag{2.3.13}$$

Note that (2.3.13) implies that

$$\text{rel int } \partial^{\uparrow}\Gamma_i \subset \Gamma_i^{\uparrow} \tag{2.3.14}$$

and

$$\Gamma_\ell \cap \Gamma_i^{\uparrow} = \emptyset \text{ for } \ell < i. \tag{2.3.15}$$

We will sometimes have to refer to the set of all *upflow successors* $\mathcal{G}_i^{\uparrow}$ or *downflow predecessors* $\mathcal{G}_i^{\downarrow}$ of Γ_i given by

$$\mathcal{G}_i^{\uparrow} := \{\Gamma_j : j \geq i\}, \quad \mathcal{G}_i^{\downarrow} := \{\Gamma_j : j \leq i\} \tag{2.3.16}$$

and their respective associated open domains

$$\Gamma_i^{\uparrow\uparrow} := \operatorname{int} \bigcup \left\{\overline{\Gamma}' : \Gamma' \in \mathcal{G}_i^{\uparrow}\right\}, \quad \Gamma_i^{\downarrow\downarrow} := \operatorname{int} \bigcup \left\{\overline{\Gamma}' : \Gamma' \in \mathcal{G}_i^{\downarrow}\right\}. \tag{2.3.17}$$

Hence aside from the list of parametric mappings the information that will ultimately be needed for the characterization of function spaces on Γ and for the subsequent construction of wavelets are the neighborhood relations encoded by $\mathcal{G}^{\uparrow}$.

2.4. A family of projections. The component spaces on the right-hand side of (2.2.12) will ultimately be identified as ranges of certain projectors. In contrast to [9] we will give an explicit (yet recursive) definition of these projectors and verify then their relevant properties.

To this end, let for $\Gamma' \subset \Gamma$ the characteristic function of Γ' be denoted by $\chi_{\Gamma'}$, i.e., $\chi_{\Gamma'}(x) = 1, x \in \Gamma'$ and $\chi_{\Gamma'}(x) = 0, x \notin \Gamma'$. Thus for any $v \in L_0(\Gamma)$, the expression $\chi_{\Gamma'} v$ means that v is first restricted to Γ' and then extended by zero to $\Gamma \backslash \Gamma'$. Likewise, with a slight abuse of notation we will also write $\chi_{\Gamma'} v$ to mean the trivial extension by zero of $v \in L_0(\Gamma')$ to all of Γ even though v may have been a priorily defined only on Γ'. The restriction of v to Γ' will be denoted by $v\,|_{\Gamma'}$.

The main ingredient for the construction of the above mentioned projections will be linear *extension* operators E_i from $L_0(\Gamma_i)$ to $L_0(\Gamma_i^{\uparrow})$, i.e.,

$$(E_i v)\,|_{\Gamma_i} = v\,|_{\Gamma_i}, \tag{2.4.1}$$

whose particular properties will be specified later.

Given such E_i we define next a family $\mathcal{P}^{\uparrow}$ of mappings P_i from $L_0(\Gamma)$ into $L_0(\Gamma)$ associated with the flow $\mathcal{G}^{\uparrow}$. For $i = 1$, let

$$P_1 v := \chi_{\Gamma_1^{\uparrow}} E_1\,(v\,|_{\Gamma_1}) \tag{2.4.2}$$

as well as

$$P_i v := \chi_{\Gamma_i^{\uparrow}} E_i \left(\left(v - \sum_{j<i} P_j v \right) |_{\Gamma_i} \right), \quad i = 2, \dots, N. \tag{2.4.3}$$

In the following we will use the form (2.4.3) also for $i = 1$, where, of course, it is understood that the sum $\sum_{j<i} P_j v$ is then vacuous and thus ignored. Clearly, each P_i depends only on a few predecessors, namely, by (2.3.15), one has

$$P_i v = \chi_{\Gamma_i^{\uparrow}} E_i \left(\left(v - \sum_{\Gamma_j \in \mathcal{N}_i^{\downarrow}} P_j v \right) |_{\Gamma_i} \right). \tag{2.4.4}$$

Of course, by definition (2.4.3), one has

$$\operatorname{supp} P_i v \subseteq \overline{\Gamma_i^{\uparrow}}. \tag{2.4.5}$$

The adjoints of the operators E_i, P_i are denoted by E_i^*, P_i^*, respectively. More precisely, one has for any $u, v \in L_2(\Gamma)$

$$\langle E_i v, w \rangle_{\Gamma_i^{\uparrow}} = \langle v, E_i^* w \rangle_{\Gamma_i}, \quad \langle P_j u, v \rangle_{\Gamma} = \langle u, P_j^* v \rangle_{\Gamma},$$

which, of course, indicates that E_i^* is a *restriction* operator.

THEOREM 2.1. *Let $\mathcal{P}^{\uparrow}$ be defined as above. The mappings*

$$Tv := ((P_i v)\,|_{\Gamma_i})_{i=1}^N, \qquad Vv := ((P_i^* v)\,|_{\Gamma_i})_{i=1}^N \tag{2.4.6}$$

are linear isomorphisms from $L_0(\Gamma)$ onto $\prod_{j=1}^N L_0(\Gamma_i)$ whose inverses are given by

$$S(v_i)_{i=1}^N = \sum_{i=1}^N P_i \chi_{\Gamma_i} v_i, \qquad U(v_i)_{i=1}^N = \sum_{i=1}^N P_i^* \chi_{\Gamma_i} v_i, \tag{2.4.7}$$

respectively.

Proof. The proof hinges on the properties of the P_i listed in the two subsequent propositions. While the development in [9] aimed at proving the existence of projectors P_i with these properties it remains to verify here that the P_i defined in (2.4.3) indeed have these properties.

PROPOSITION 2.4.1. *The P_i defined above have the following properties:*

(i) *One has*

$$P_iP_j = \delta_{i,j}P_i, \quad 1 \leq i,\, j \leq N. \tag{2.4.8}$$

(ii) *For any $v \in L_0(\Gamma)$, one has*

$$v = \sum_{j=1}^{N} P_j v. \tag{2.4.9}$$

(iii) *One has*

$$\chi_{\Gamma_i} P_j v = P_i \chi_{\Gamma_j} v = 0, \quad 1 \leq i < j \leq N. \tag{2.4.10}$$

Since these verifications are elementary but after all helpful to keep the present approach self-contained, they will be deferred to Appendix A.

We will have to deal with the adjoints P_i^* as well. To this end, note that

$$\langle P_i v, w\rangle_\Gamma = \left\langle \left(v - \sum_{j<i} P_j v\right) |_{\Gamma_i}, E_i^*(w\,|_{\Gamma_i^\uparrow})\right\rangle_\Gamma = \left\langle v - \sum_{j<i} P_j v, \chi_{\Gamma_i} E_i^*(w\,|_{\Gamma_i^\uparrow})\right\rangle_\Gamma,$$

i.e.,

$$P_i^* w = \left(I - \sum_{j<i} P_j^*\right) \chi_{\Gamma_i} E_i^*(w\,|_{\Gamma_i^\uparrow}). \tag{2.4.11}$$

It is now easy to establish the following analogous statements for the adjoints.

Proposition 2.4.2. *The adjoints P_i^* have analogous properties, i.e.,*

$$P_i^* P_j^* = \delta_{i,j} P_i^*, \tag{2.4.12}$$

$$v = \sum_{j=1}^{N} P_j^* v, \tag{2.4.13}$$

and

$$\chi_{\Gamma_i} P_j^* v = P_i^* \chi_{\Gamma_j} v = 0, \quad 1 \leq j < i \leq N. \tag{2.4.14}$$

Proof. Relations (2.4.12) and (2.4.13) follow immediately from (2.4.8) and (2.4.9) by duality. Likewise, by (2.4.10), one has for $j < i$

$$0 = \left\langle \chi_{\Gamma_j} P_i v, w\right\rangle_\Gamma = \left\langle v, P_i^* \chi_{\Gamma_j} w\right\rangle_\Gamma$$

and also

$$0 = \langle P_j \chi_{\Gamma_i} v, w\rangle_\Gamma = \left\langle v, \chi_{\Gamma_i} P_j^* w\right\rangle_\Gamma. \qquad \square$$

Thanks to properties established in Propositions 2.4.1 and 2.4.2 the proof of Theorem 2.1 is essentially the same as in [9]. For the convenience of the reader and since the result is of central importance, we sketch the argument. By definition of T and S one obtains

$$S(Tv) = \sum_{i=1}^{N} P_i \chi_{\Gamma_i}(P_i v)\,|_{\Gamma_i} = \sum_{i=1}^{N} P_i \chi_{\Gamma_i} P_i v. \tag{2.4.15}$$

By (2.4.8), one has

$$P_i v = P_i^2 v = P_i \left(\sum_{j=1}^{N} \chi_{\Gamma_j} \right) P_i v = P_i \left(\sum_{j=1}^{i} \chi_{\Gamma_j} \right) P_i v = P_i \chi_{\Gamma_i} P_i, \tag{2.4.16}$$

where we have used (2.4.10) in the last two steps. Combining (2.4.16) with (2.4.15) and bearing (2.4.9) in mind yields

$$S(Tv) = \sum_{i=1}^{N} P_i v = v,$$

which proves that $T^{-1} = S$.

As for the remaining part of the claim, one can use the same arguments based on Proposition 2.4.2. Alternatively, one can exploit that obviously

$$U = T^*, \quad V = S^*. \tag{2.4.17}$$

Thus, by the previous argument

$$V = S^* = (T^{-1})^* = (T^*)^{-1} = U^{-1},$$

which completes the proof of Theorem 2.1. □

2.5. Topological isomorphisms. The properties of the mappings T, V and their inverses are so far purely algebraic. We will show next that their *topological* properties are completely determined by the continuity properties of the extension operators E_i which will be described next.

To this end, let us first fix some notation and conventions that will be used throughout the remainder of the paper. As above $\mathcal{F}$ will always stand for a Besov function space of the type $\mathcal{F} = B_q^s(L_p)$, $s > 0$ ($s \geq 0$ when $p = q = 2$), where s is bounded from above by some s_Γ depending on Γ; see, e.g., [4, 26, 46] for the precise definition of these Besov spaces. Occasionally we will refer to the specific regularity index by writing $\mathcal{F}_s$ or $\mathcal{F}_\tau$ for $\tau \leq s$. Throughout the following $s \leq s_\Gamma$ will be fixed describing the range in which subsequent topological properties will be considered. For us it is important that $B_2^s(L_2) = H_2^s =: H^s$, i.e., for $p = q = 2$ the Besov space coincides with the standard Sobolev space relative to the L_2 up to norm equivalence.

Many of the results will actually hold for $p, q \in (0, \infty]$. When duality enters one often has to restrict the discussion to $p, q \geq 1$. In this context, given $\mathcal{F} = B_q^\tau(L_p)$, the corresponding scale of spaces $\tilde{\mathcal{F}}_\tau = B_{q'}^\tau(L_{p'})$ with *adjoint* indices p', q' satisfying $\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1$ comes into play.

Moreover, the Besov spaces are *interpolation spaces* with respect to real interpolation [4, 46],

$$\mathcal{F}_\tau = (H_p^t, L_p)_{\tau/t,q} = (\mathcal{F}_s, \mathcal{F}_r)_{\theta,q}, \quad \tau = \theta s + (1-\theta) r, \tag{2.5.1}$$

where H_p^t denotes the usual Sobolev (respectively, Bessel potential) spaces with respect to L_p; see, e.g., [4, 46].

Recall that by definition

$$\|v\|_{\mathcal{F}(\Gamma')} \lesssim \|v\|_{\mathcal{F}(\Gamma'')} \tag{2.5.2}$$

holds for any domains $\Gamma' \subseteq \Gamma'' \subseteq \Gamma$. Now define the spaces

$$\mathcal{F}(\Gamma_i)^{\uparrow} \quad (\mathcal{F}(\Gamma_i)^{\downarrow}) := \left\{ v \in \mathcal{F}(\Gamma_i) \,:\, \chi_{\Gamma_i} v \in \mathcal{F}(\Gamma_i^{\uparrow}) \quad (\mathcal{F}(\Gamma_i^{\downarrow})) \right\}, \tag{2.5.3}$$

endowed with the norms

$$\|v\|_{\mathcal{F}(\Gamma_i)^{\uparrow}} := \|\chi_{\Gamma_i} v\|_{\mathcal{F}(\Gamma_i^{\uparrow})}\,, \qquad \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} := \|\chi_{\Gamma_i} v\|_{\mathcal{F}(\Gamma_i^{\downarrow})}\,. \tag{2.5.4}$$

Suppose that $v \in \mathcal{F}(\Gamma_i^{\uparrow})$ is the strong limit of a sequence $v_l \in \mathcal{F}(\Gamma_i)^{\uparrow}$. Since for every C^∞-function φ with compact support in $\Gamma_i^{\uparrow} \setminus \Gamma_i$ one has

$$|\langle v, \varphi\rangle_{\Gamma_i^{\uparrow}}| = |\langle v - v_l, \varphi\rangle_{\Gamma_i^{\uparrow}}| \le \|v - v_l\|_{\mathcal{F}(\Gamma_i^{\uparrow})} \|\varphi\|_{\mathcal{F}^*(\Gamma_i^{\uparrow})} \to 0,$$

we see that $\chi_{\Gamma_i} v = v$ which confirms the following fact.

REMARK 2.5.1. *$\mathcal{F}(\Gamma_i)^{\uparrow}$, $\mathcal{F}(\Gamma_i)^{\downarrow}$ are closed subspaces of $\mathcal{F}(\Gamma_i)$ with respect to the norms defined in* (2.5.4).

Since $\mathcal{F}(\Gamma_i)^{\uparrow}$ consists of exactly those elements in $\mathcal{F}(\Gamma_i)$ whose trivial extension by zero across the outflow boundary $\partial^{\uparrow}\Gamma_i$ belongs to the corresponding space $\mathcal{F}(\Gamma_i^{\uparrow})$ for the extended domain, the elements in $\mathcal{F}(\Gamma_i)^{\uparrow}$ are characterized by the fact that their trace vanishes on the outflow boundary $\partial^{\uparrow}\Gamma_i$ in a certain sense. More precisely, although we will not make explicit use of it, we recall the following fact; see, e.g., [46, Section 2.10.2, Theorem 1] or [29, Theorem 1.4.5.2].

REMARK 2.5.2. *$\mathcal{F}_\tau(\Gamma_i)^{\uparrow}$ agrees with the closure in $\|\cdot\|_{\mathcal{F}_\tau(\Gamma_i)}$ of all smooth functions whose support is contained in $\overline{\Gamma}_i$ but does not intersect $\partial^{\uparrow}\Gamma_i$ provided that $\tau + 1/p$ is* not an integer. *The same holds for $\mathcal{F}(\Gamma_i)^{\downarrow}$ relative to the reverse flow. Moreover one has*

$$\|\cdot\|_{\mathcal{F}_\tau(\Gamma_i)^{\uparrow}} = \|\cdot\|_{\mathcal{F}_\tau(\Gamma_i)^{\downarrow}} = \|\cdot\|_{\mathcal{F}_\tau(\Gamma_i)} \quad \textit{if} \;\; \tau + 1/p \notin \mathbb{N}. \tag{2.5.5}$$

Throughout the remaining part of this section we will assume the following properties of the extension operators E_i:

ASSUMPTION A.

- *Localness*:

$$\chi_{\Gamma_i^{\uparrow}} E_i v \in \mathcal{F}(\Gamma_i^{\uparrow\uparrow}), \quad v \in \mathcal{F}(\Gamma_i), \;\; i = 1, \dots, N. \tag{2.5.6}$$

- *Continuity of E_i*:

$$\|E_i v\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} \lesssim \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}}\,, \quad v \in \mathcal{F}(\Gamma_i)^{\downarrow}, \;\; i = 1, \dots, N. \tag{2.5.7}$$

- *Continuity of E_i^**:

$$\|E_i^* v\|_{\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}} \lesssim \|v\|_{\tilde{\mathcal{F}}(\Gamma_i^{\uparrow})}\,, \quad v \in \tilde{\mathcal{F}}(\Gamma_i^{\uparrow}), \;\; i = 1, \dots, N. \tag{2.5.8}$$

One could actually require *different* regularity parameters $s', \tilde{s}'$ in (2.5.7) and (2.5.8), respectively, which will naturally come up in section 4.3 below. Since this will not be essential for the subsequent analysis, we dispense here with this complication of notation.

Remark 2.5.3. *We will always assume that Assumption A holds, in particular, for the L_p-norm corresponding to $\mathcal{F} = \mathcal{F}_s$. The spaces $\mathcal{F}(\Gamma_i)^{\uparrow}$ are also* stable *under interpolation*

$$\mathcal{F}_\tau(\Gamma_i)^{\uparrow} = (\mathcal{F}_s(\Gamma_i)^{\uparrow}, L_p)_{\tau/s,q} \tag{2.5.9}$$

(and likewise for $\uparrow$ replaced by $\downarrow$), see, e.g., [38], *so that Assumption A will actually hold for the full range between* 0 *and* s.

Next we will comment on the dual spaces for spaces of the type $\mathcal{F}(\Gamma_i)^{\uparrow}$.

Remark 2.5.4. *Suppose that $\Gamma' \subset \Gamma$. Then $w \in \mathcal{F}^*(\Gamma')$, the dual of $\mathcal{F}(\Gamma')$ (the space of bounded linear functionals on $\mathcal{F}(\Gamma')$), iff $\chi_{\Gamma'} w \in \mathcal{F}^*(\Gamma)$ and*

$$\|\chi_{\Gamma'} w\|_{\mathcal{F}^*(\Gamma)} \sim \|w\|_{\mathcal{F}^*(\Gamma')}. \tag{2.5.10}$$

Now consider the closed subspace

$$\mathcal{F}_i := \{v \in \mathcal{F}(\Gamma) : \chi_{\Gamma_i^{\uparrow\uparrow}} v \in \mathcal{F}(\Gamma)\}, \quad \|v\|_{\mathcal{F}_i} := \|\chi_{\Gamma_i^{\uparrow\uparrow}} v\|_{\mathcal{F}(\Gamma)}$$

of $\mathcal{F}(\Gamma)$, consisting of those elements in $\mathcal{F}(\Gamma_i^{\uparrow\uparrow})$ whose trivial extension by zero belongs to $\mathcal{F}(\Gamma)$. The same reasoning as above yields the following fact.

Remark 2.5.5. $w \in \big(\mathcal{F}(\Gamma_i)^{\downarrow}\big)^*$ *iff* $\chi_{\Gamma_i} w \in \mathcal{F}_i^*$ *and*

$$\|\chi_{\Gamma_i} w\|_{\mathcal{F}_i^*} \sim \|w\|_{(\mathcal{F}(\Gamma_i)^{\downarrow})^*}. \tag{2.5.11}$$

Thus $\big(\mathcal{F}(\Gamma_i)^{\downarrow}\big)^*$ consists of those functionals whose trivial extension by zero *only* in the outflow direction remains continuous. In this sense it is justified to set

$$\big(\mathcal{F}(\Gamma_i)^{\downarrow}\big)^* = \mathcal{F}^*(\Gamma_i)^{\uparrow}, \quad \big(\mathcal{F}(\Gamma_i)^{\uparrow}\big)^* = \mathcal{F}^*(\Gamma_i)^{\downarrow}. \tag{2.5.12}$$

We are now ready to state the first main result in this section.

Theorem 2.2. *Under the assumptions* (2.5.7) *and* (2.5.8) *the projectors P_j defined by* (2.4.2) *and* (2.4.3) *satisfy*

$$\begin{aligned} (P_i \chi_{\Gamma_i^{\uparrow\uparrow}} v)\,|_{\Gamma_i} &= v\,|_{\Gamma_i} \quad \forall \;\; v \in \mathcal{F}(\Gamma), \;\; v\,|_{\Gamma_i} \in \mathcal{F}(\Gamma_i)^{\downarrow}, \\ (P_i^* \chi_{\Gamma_i^{\downarrow\downarrow}} v)\,|_{\Gamma_i} &= v\,|_{\Gamma_i} \quad \forall \;\; v \in \tilde{\mathcal{F}}(\Gamma), \;\; v\,|_{\Gamma_i} \in \tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}, \end{aligned} \tag{2.5.13}$$

and their restriction to Γ_i induces topological isomorphisms from $P_i(\mathcal{F}(\Gamma))$, $P_i^*(\tilde{\mathcal{F}}(\Gamma))$ onto $\mathcal{F}(\Gamma_i)^{\downarrow}$, $\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}$, respectively, i.e.,

$$(P_i(\mathcal{F}(\Gamma)))\,|_{\Gamma_i} = \mathcal{F}(\Gamma_i)^{\downarrow}, \quad (P_i^*(\tilde{\mathcal{F}}(\Gamma)))\,|_{\Gamma_i} = \tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}, \tag{2.5.14}$$

where

$$\|(P_i v)\,|_{\Gamma_i}\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} \;\lesssim\; \|v\|_{\mathcal{F}(\Gamma)}\,, \quad \|(P_i^* v)\,|_{\Gamma_i}\|_{\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}} \;\lesssim\; \|v\|_{\tilde{\mathcal{F}}(\Gamma)}\,. \tag{2.5.15}$$

Moreover, the P_i as well as their adjoints P_i^* are continuous in $\mathcal{F}(\Gamma)$, $\tilde{\mathcal{F}}(\Gamma)$, respectively, i.e., one has

$$\|P_i v\|_{\mathcal{F}(\Gamma)} \lesssim \|v\|_{\mathcal{F}(\Gamma)}, \quad v \in \mathcal{F}(\Gamma), \quad \|P_i^* v\|_{\tilde{\mathcal{F}}(\Gamma)} \lesssim \|v\|_{\tilde{\mathcal{F}}(\Gamma)}, \quad v \in \tilde{\mathcal{F}}(\Gamma). \tag{2.5.16}$$

An analogous result has been already established in [9]. However, since for the anticipated applications it will be important to apply the above characterizations with the particular projectors P_j defined in (2.4.3) and since in this form they do not seem to be readily identified with the corresponding quantities occurring in [9], we will include a self-contained proof which will be deferred to Appendix B.

By Theorem 2.2, the mappings T and V from (2.4.6) and (2.4.7) actually induce now mappings

$$T : \mathcal{F}(\Gamma) \to \Pi\mathcal{F}^{\downarrow} := \prod_{i=1}^{N} \mathcal{F}(\Gamma_i)^{\downarrow}, \quad V : \tilde{\mathcal{F}}(\Gamma) \to \Pi\tilde{\mathcal{F}}^{\uparrow} := \prod_{i=1}^{N} \tilde{\mathcal{F}}(\Gamma_i)^{\uparrow},$$

which will turn out to be *topological isomorphisms.* The main result of this section reads as follows.

THEOREM 2.3. *One has*

$$\mathcal{F}(\Gamma) \cong \prod_{i=1}^{N} \mathcal{F}(\Gamma_i)^{\downarrow}, \quad \tilde{\mathcal{F}}(\Gamma) \cong \prod_{i=1}^{N} \tilde{\mathcal{F}}(\Gamma_i)^{\uparrow},$$

and defining

$$|||v|||_{\mathcal{F}^{\downarrow}} := \left(\sum_{i=1}^{N} \|(P_i v)\ |_{\Gamma_i}\|^2_{\mathcal{F}(\Gamma_i)^{\downarrow}} \right)^{\frac{1}{2}}, \quad |||v|||_{\tilde{\mathcal{F}}^{\uparrow}} := \left(\sum_{i=1}^{N} \|(P_i^* v)\ |_{\Gamma_i}\|^2_{\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}} \right)^{\frac{1}{2}}, \tag{2.5.17}$$

the norms $\|\cdot\|_{\tilde{\mathcal{F}}(\Gamma)}$, $\|\cdot\|_{\mathcal{F}(\Gamma)}$ and $|||\cdot|||_{\tilde{\mathcal{F}}^{\uparrow}}$, $|||\cdot|||_{\mathcal{F}^{\downarrow}}$, respectively, are equivalent, i.e.,

$$\|v\|_{\tilde{\mathcal{F}}(\Gamma)} \sim |||v|||_{\tilde{\mathcal{F}}^{\uparrow}}, \quad v \in \tilde{\mathcal{F}}(\Gamma), \qquad \|v\|_{\mathcal{F}(\Gamma)} \sim |||v|||_{\mathcal{F}^{\downarrow}}, \quad v \in \mathcal{F}(\Gamma). \tag{2.5.18}$$

Moreover, V and T extend to isomorphisms from $\mathcal{F}^*(\Gamma)$, $\tilde{\mathcal{F}}^*(\Gamma)$ onto the spaces $\Pi\mathcal{F}^*_{\uparrow} := \prod_{i=1}^{N} \mathcal{F}^*(\Gamma_i)^{\uparrow}$ and $\Pi\tilde{\mathcal{F}}^*_{\downarrow} := \prod_{i=1}^{N} \tilde{\mathcal{F}}^*(\Gamma_i)^{\downarrow}$, respectively, and

$$\|v\|_{\tilde{\mathcal{F}}^*(\Gamma)} \sim \left(\sum_{j=1}^{N} \left\|(P_j v)\ |_{\Gamma_j}\right\|^2_{\tilde{\mathcal{F}}^*(\Gamma_j)^{\downarrow}} \right)^{\frac{1}{2}}, \quad \|v\|_{\mathcal{F}^*(\Gamma)} \sim \left(\sum_{j=1}^{N} \left\|(P_j^* v)\ |_{\Gamma_j}\right\|^2_{\mathcal{F}^*(\Gamma_j)^{\uparrow}} \right)^{\frac{1}{2}}. \tag{2.5.19}$$

Proof. With Theorem 2.2 and the other prerequisites at hand the proof follows exactly the reasoning of [9]. In fact, the availability of algebraic inverses combined with the open mapping theorem yields (2.5.18). As for the dual spaces, the first part of the theorem ensures that $V : \tilde{\mathcal{F}}(\Gamma) \to \Pi\tilde{\mathcal{F}}^{\uparrow}$ and hence $V^* : (\Pi\tilde{\mathcal{F}}^{\uparrow})^* \to \tilde{\mathcal{F}}^*(\Gamma)$ are isomorphisms. By denseness of the involved spaces in their dual versions and noting that $U = V^{-1}$ the mapping U^* extends to an isomorphism from $\tilde{\mathcal{F}}^*(\Gamma)$ onto $(\Pi\tilde{\mathcal{F}}^{\uparrow})^* = \Pi\tilde{\mathcal{F}}^*_{\downarrow}$, where we have used (2.5.12) in the last step. The reasoning for $v = S^*$ is analogous. The assertion follows now from (2.4.17). □

In view of Theorems 2.2 and 2.3, the construction of wavelet bases on Γ satisfying (I) in section 1.1 amounts to the following two tasks:

(a) The construction of wavelet bases for the component spaces $\tilde{\mathcal{F}}(\Gamma_i)^\uparrow$ or $\mathcal{F}(\Gamma_i)^\downarrow$ with respective properties (I).

(b) The (practical) realization of extension operators E_i satisfying Assumption A.

We will first address (a) which incidentally will be crucial for dealing with (b) as well.

3. Discrete norm equivalences. We adhere to the above notation and recall that the parametric mappings κ_i will always be assumed to be regular parametrizations of any degree of smoothness required at each instance.

Our goal is to establish isomorphisms from $\mathcal{F}(\Gamma)$ and $\tilde{\mathcal{F}}(\Gamma)$ onto *sequence spaces* which consist of expansion sequences of the elements of $\mathcal{F}(\Gamma)$ and $\tilde{\mathcal{F}}(\Gamma)$ with respect to certain wavelet bases. In principle, such isomorphisms have been already constructed in [9]. However, the present construction differs from that in [9] in several respects. The main point here is that the underlying bases are completely *local* and that the involved extension operators in (2.4.6) affect *only* basis functions *near* the patch boundaries.

The desired norm equivalences will be based on Theorem 2.3. Thus one has to identify first suitable bases for the component spaces $\mathcal{F}(\Gamma_i)^\downarrow$, $\tilde{\mathcal{F}}(\Gamma_i)^\uparrow$. Due to the regularity of the mappings κ_i these spaces can be essentially identified with corresponding smoothness spaces over the parameter domain $\Box$. Thus the construction will be divided into the following steps:

(i) Employ suitable biorthogonal wavelet bases over $\Box$ which give rise there to discrete norm equivalences for the pull backs of the spaces $\mathcal{F}(\Gamma_i)^\downarrow$ and their duals $\mathcal{F}^*(\Gamma_i)^\uparrow$.

(ii) Lift these bases parametrically to the patches Γ_i and verify the validity of corresponding norm equivalences.

(iii) Use Theorem 2.3 to construct bases on Γ along with corresponding isomorphisms.

We emphasize already at this point that the construction of bases on Γ is primarily a conceptual issue. Actual computations will be seen later to involve only bases of the component spaces $\tilde{\mathcal{F}}(\Gamma_i)^\uparrow, \mathcal{F}(\Gamma_i)^\downarrow$ or better yet bases on the parameter domain $\Box$.

3.1. Wavelet bases on $\Box$. Wavelet bases on $\Box = (0,1)^n$ are conveniently constructed with the aid of *tensor products* of wavelet bases on $[0,1]$. The essential property of such pairs of biorthogonal wavelet bases on $[0,1]$ will be seen to be certain *complementary boundary conditions*. Such bases have been constructed in [22] and this section is devoted to briefly summarizing the relevant properties needed in the subsequent development.

To this end, we will consistently use the following notation:

$$\Psi = \{\psi_\lambda : \lambda \in \nabla\}, \quad \tilde{\Psi} = \{\tilde{\psi}_\lambda : \lambda \in \nabla\}$$

will always denote a pair of biorthogonal wavelet bases (relative to some inner product to be specified at each instance). The indices $\lambda \in \nabla$ are to encode the level $|\lambda|$ of resolution of ψ_λ (typically representing a mesh size of order $2^{-|\lambda|}$) as well as the location and type of the wavelet ψ_λ. For instance, the wavelets ψ_λ supported in the interior of the respective domain will have the form $\psi_\lambda = 2^{jn/2}\psi_e(2^j \cdot -k)$ with $e \in \{0,1\}^n \setminus \{\mathbf{0}\}$, $k \in \mathbb{Z}^n$, i.e., $\lambda = 2^{-j}(k + \frac{e}{2})$. By $\Psi_j = \{\psi_\lambda : |\lambda| = j\}$ we mean all wavelets in Ψ on level j. There will always be some coarsest level $j_0 \in \mathbb{N}$ for which in addition to the complement basis Ψ_{j_0} we also need a basis $\Phi_{j_0} = \{\phi_\lambda : \lambda \in \Delta_+\}$ whose elements are of *scaling function type* (adapted to the domain) containing polynomials up to some fixed order d (degree $d-1$). Thus the whole index set ∇ is split into $\nabla = \Delta_+ \cup \nabla_-$, where Δ_+ refers to the coarse level (polynomial) part and ∇_- represents the “true” wavelets. Such bases will be considered for various domains Ω such as $\square$, Γ_i, Γ which will be indicated by corresponding superscripts. Likewise, the reference of index sets to respective bases will be indicated by superscripts. For instance, ∇^i refers to a basis on Γ_i.

Furthermore, we will consistently make use of the following conventions. Any collection Φ of functions will be viewed as a (column) vector whose components are the elements of Φ with respect to some fixed but unspecified order. Thus

$$\mathbf{c}^T\Phi = \sum_{\phi\in\Phi} c_\phi \phi$$

denotes corresponding linear combinations where coefficient vectors are kept in boldface. Likewise given any dual form $\langle\cdot,\cdot\rangle$

$$\langle\Phi,\Theta\rangle = (\langle\phi,\theta\rangle)_{\phi\in\Phi,\theta\in\Theta}$$

is a matrix. In particular, for any function v the expressions $\langle v,\Phi\rangle$, $\langle\Phi,v\rangle$ are row and column vectors, respectively. By $\mathbf{I}$ we will denote the identity matrix whose dimensionality will always be clear from the context.

According to the nature of the spaces $\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}$ and $\mathcal{F}(\Gamma_i)^{\downarrow}$, we have to deal with analogous spaces on $\square$ whose elements admit zero extensions across all possible unions of facets of $\square$. To organize this we follow essentially the notation in [9] and describe first how to extend the cube $\square$ across such faces in order to define suitable preimages of the sets $\Gamma_i^{\uparrow}$, $\Gamma_i^{\downarrow}$. To this end, consider first the univariate case and let for $Z \subseteq \{0,1\}$

$$[0,1]_Z := \begin{cases} [0,1] & \text{if } Z=\emptyset, \\ [-1,1] & \text{if } Z=\{0\}, \\ [0,2] & \text{if } Z=\{1\}, \\ [-1,2] & \text{if } Z=\{0,1\}. \end{cases} \tag{3.1.1}$$

For $n>1$, we simply set $\mathbf{Z} = Z_1\times\cdots\times Z_n \subseteq \{0,1\}^n$ and define

$$\square_{\mathbf{Z}} = [0,1]_{Z_1}\times\cdots\times[0,1]_{Z_n}. \tag{3.1.2}$$

Then

$$\partial_{\mathbf{Z}}\square := \bigcup_{\ell=1}^{n} [0,1]^{\ell-1}\times Z_\ell\times[0,1]^{n-\ell} \tag{3.1.3}$$

is the union of those faces of $\square$ across which $\square$ has been extended to $\square_{\mathbf{Z}}$. We will consistently use the notation

$$\tilde{Z} := \{0,1\}\setminus Z, \qquad \tilde{\mathbf{Z}} := \{0,1\}^n\setminus\mathbf{Z}, \tag{3.1.4}$$

so that

$$\partial_{\tilde{\mathbf{Z}}}\square = \partial\square\setminus\partial_{\mathbf{Z}}\square, \tag{3.1.5}$$

where as usual $\partial\square$ denotes the boundary of $\square$. In complete analogy to (2.5.3) we define the spaces

$$\mathcal{F}_\tau(\square)_{\mathbf{Z}} := \{v\in\mathcal{F}_\tau(\square) : \chi_\square v\in\mathcal{F}_\tau(\square_{\mathbf{Z}})\},$$

endowed with the norm

$$\|v\|_{\mathcal{F}_\tau(\Box)_{\mathbf{Z}}} := \|\chi_\Box v\|_{\mathcal{F}_\tau(\Box_{\mathbf{Z}})}. \tag{3.1.6}$$

Throughout the rest of this section we will make use of the fact that for any pair of regularity parameters $\gamma = \gamma(p) > 0$, $\tilde{\gamma} = \tilde{\gamma}(p') > 0$ and for each choice of $\mathbf{Z}$ certain pairs of *biorthogonal wavelet bases*

$$\Psi^{\mathbf{Z}} = \{\psi_\lambda^{\mathbf{Z}} : \lambda \in \nabla^{\mathbf{Z}}\}, \quad \tilde{\Psi}^{\tilde{\mathbf{Z}}} = \{\tilde{\psi}_\lambda^{\tilde{\mathbf{Z}}} : \lambda \in \nabla^{\mathbf{Z}}\}$$

(identifying their respective index sets) on $\Box$ have been constructed in [22] which will be said to have *type* $\mathbf{Z}, \tilde{\mathbf{Z}}$, respectively, such that for $d, \tilde{d} \in \mathbb{N}$, $d + \tilde{d} \in 2\mathbb{N}$, sufficiently large, the following properties hold; see Theorem 3.2.1 in [22].

Properties B.

(i) *The collections* $\Psi^{\mathbf{Z}}, \tilde{\Psi}^{\tilde{\mathbf{Z}}}$ *are biorthogonal, i.e.,*

$$\langle \Psi^{\mathbf{Z}}, \tilde{\Psi}^{\tilde{\mathbf{Z}}} \rangle_\Box = \mathbf{I}. \tag{3.1.7}$$

Moreover, one has

$$\Psi^{\mathbf{Z}} \subset \mathcal{F}_\tau(\Box) \quad \textit{for} \quad \tau < \gamma, \quad \tilde{\Psi}^{\tilde{\mathbf{Z}}} \subset \tilde{\mathcal{F}}_\tau(\Box) \quad \textit{for} \quad \tau < \tilde{\gamma}. \tag{3.1.8}$$

In addition, the primal and dual wavelets satisfy certain (complementary) homogeneous boundary conditions *in the following sense. For some fixed* $0 < s' < \gamma$, $0 < \tilde{s}' < \tilde{\gamma}$ *one has*

$$\Psi^{\mathbf{Z}} \subset \mathcal{F}_\tau(\Box)_{\mathbf{Z}}, \quad \tau \leq s', \quad \tilde{\Psi}^{\tilde{\mathbf{Z}}} \subset \tilde{\mathcal{F}}_\tau(\Box)_{\tilde{\mathbf{Z}}}, \quad \tau \leq \tilde{s}'. \tag{3.1.9}$$

(ii) *The bases* $\Psi^{\mathbf{Z}}$, $\tilde{\Psi}^{\tilde{\mathbf{Z}}}$ *are local, i.e., for* $\Omega_\lambda^\Box := \operatorname{supp} \psi_\lambda^{\mathbf{Z}}$, $\tilde{\Omega}_\lambda^\Box := \operatorname{supp} \tilde{\psi}_\lambda^{\tilde{\mathbf{Z}}}$ *one has*

$$\operatorname{diam} \Omega_\lambda^\Box, \quad \operatorname{diam} \tilde{\Omega}_\lambda^\Box \sim 2^{-|\lambda|}. \tag{3.1.10}$$

(iii) *For* $s', \tilde{s}'$ *as above one has*

$$\Pi_d(\Box) \cap \mathcal{F}_{s'}(\Box)_{\mathbf{Z}} \subset \operatorname{span}\{\psi_\lambda^{\mathbf{Z}} : \lambda \in \Delta_+^{\mathbf{Z}}\},$$

and

$$\Pi_{\tilde{d}}(\Box) \cap \tilde{\mathcal{F}}_{\tilde{s}'}(\Box)_{\tilde{\mathbf{Z}}} \subset \operatorname{span}\{\tilde{\psi}_\lambda^{\tilde{\mathbf{Z}}} : \lambda \in \Delta_+^{\mathbf{Z}}\},$$

where $\Pi_d(\Omega)$ *denotes the space of all polynomials of order* d *on* Ω. *Moreover, for some* $a > 0$ *and any* $\Omega \subset \Box$ *with* $\operatorname{dist}(\partial_{\mathbf{Z}}\Box, \partial\Omega) > a2^{-j}$ *one has*

$$\Pi_d(\Omega) \subseteq \operatorname{span}\{\psi_\lambda^{\mathbf{Z}} : \lambda \in \nabla^{\mathbf{Z}}, |\lambda| < j\} =: S_{j,\mathbf{Z}}.$$

Likewise, when $\operatorname{dist}(\partial_{\tilde{\mathbf{Z}}}\Box, \partial\Omega) > a2^{-j}$ *one has*

$$\Pi_{\tilde{d}}(\Omega) \subseteq \operatorname{span}\{\tilde{\psi}_\lambda^{\tilde{\mathbf{Z}}} : \lambda \in \nabla^{\mathbf{Z}}, |\lambda| < j\} =: \tilde{S}_{j,\tilde{\mathbf{Z}}},$$

that is, away from the part of the boundary marked by $\mathbf{Z}$, $\tilde{\mathbf{Z}}$, *the primal and dual multiresolution spaces contain polynomials up to order* $d, \tilde{d}$, *respectively.*

(iv) *There exists* $\Lambda_{\mathbf{Z}}^\emptyset \subset \nabla^{\mathbf{Z}}$ *such that*

- $\nu \in \nabla^{\mathbf{Z}} \setminus \Lambda_{\mathbf{Z}}^\emptyset$ *implies that* $\operatorname{dist}(\partial_{\mathbf{Z}}\Box, \Omega_\nu^\Box) \lesssim 2^{-|\nu|}$.

- *The collection* $\{\psi_\nu^{\mathbf{Z}} : \nu \in \Lambda_{\mathbf{Z}}^{\emptyset}\}$ *can be extended to a collection* $\Psi^{\emptyset}$ *of type* $\mathbf{Z} = \emptyset$ *(having, therefore, by* (iii)*, full polynomial exactness of order* d *on all of* $\square$*). Moreover, the corresponding biorthogonal collection* $\tilde{\Psi}^{\{0,1\}^n}$ *contains* $\{\tilde{\psi}_\nu^{\tilde{\mathbf{Z}}} : \nu \in \Lambda_{\mathbf{Z}}^{\emptyset}\}$ *and belongs to* $\tilde{\mathcal{F}}_{\tilde{s}'}(\square)_{\{0,1\}^n}$*.*

Likewise, there exists $\Lambda_{\mathbf{Z}}^{\{0,1\}^n} \subset \nabla^{\mathbf{Z}}$ *such that*

- $\nu \in \nabla^{\mathbf{Z}} \setminus \Lambda_{\mathbf{Z}}^{\{0,1\}^n}$ *implies that* $\operatorname{dist}(\partial_{\tilde{\mathbf{Z}}}\square, \Omega_\nu^{\square}) \lesssim 2^{-|\nu|}$*.*
- *The collection* $\{\psi_\nu^{\mathbf{Z}} : \nu \in \Lambda_{\mathbf{Z}}^{\{0,1\}^n}\}$ *can be extended to a collection* $\Psi^{\{0,1\}^n}$ *of type* $\mathbf{Z} = \{0,1\}^n$*. Moreover, the corresponding biorthogonal collection* $\tilde{\Psi}^{\emptyset}$ *contains* $\{\tilde{\psi}_\nu^{\tilde{\mathbf{Z}}} : \nu \in \Lambda_{\mathbf{Z}}^{\{0,1\}^n}\}$ *and reproduces all polynomials in* $\Pi_{\tilde{d}}$ *on* $\square$*.*

All wavelet bases are local in the sense of (3.1.10)*.*

In particular, (iii) implies the *moment conditions*

$$\langle P, \psi_\lambda^{\mathbf{Z}}\rangle = 0,\ \ P \in \Pi_{\tilde{d}}(\square)\cap\tilde{\mathcal{F}}_{\tilde{s}'}(\square)_{\tilde{\mathbf{Z}}},\ \ \langle P, \tilde{\psi}_\lambda^{\tilde{\mathbf{Z}}}\rangle = 0,\ \ P \in \Pi_d(\square)\cap\mathcal{F}_{s'}(\square)_{\mathbf{Z}},\ \ \text{for } \lambda \in \nabla_-^{\mathbf{Z}}. \tag{3.1.11}$$

Moreover, (iv) says that away from the boundary of $\square$ the wavelets of *all* types $\mathbf{Z}$ can actually be arranged to *coincide* and that the adaptation of boundary conditions on level j affects only a margin of width 2^{-j}, which will turn out to have important implications.

Since the wavelets will ultimately be transported to the manifold Γ, we may assume without loss of generality that throughout the remainder of the paper

$$s', \tilde{s}' \le s < s_\Gamma. \tag{3.1.12}$$

In addition to the above structural properties pertaining mainly to *localization* and (local) polynomial exactness we will make essential use of the following *topological* properties from Theorems 3.3.1 and 3.4.1 in [22]. To this end, since we have to deal with spaces for which the regularity is not tied to the order of boundary conditions, consider

$$\mathcal{F}_{\tau,s'}(\square)_{\mathbf{Z}} := \begin{cases} \mathcal{F}_\tau(\square)_{\mathbf{Z}} & \text{for} \quad \tau \le s', \\ \mathcal{F}_{s'}(\square)_{\mathbf{Z}} \cap \mathcal{F}_\tau(\square) & \text{for} \quad s' < \tau < d, \\ H_p^d(\square) \cap \mathcal{F}_{s'}(\square)_{\mathbf{Z}} & \text{for} \quad \tau = d. \end{cases} \tag{3.1.13}$$

The spaces $\tilde{\mathcal{F}}_{\tau,\tilde{s}'}(\square)_{\tilde{\mathbf{Z}}}$ are defined analogously with $d, p, s, \mathbf{Z}$ replaced by $\tilde{d}, p', \tilde{s}', \tilde{\mathbf{Z}}$, respectively.

Properties C.

(i) *The following Jackson inequalities hold for the multiresolution spaces* $S_{j,\mathbf{Z}}, \tilde{S}_{j,\tilde{\mathbf{Z}}}$ *defined in Properties B* (iii)*. One has*

$$\inf_{v_j \in S_{j,\mathbf{Z}}} \|v - v_j\|_{\mathcal{F}_{t,s'}(\square)_{\mathbf{Z}}} \lesssim 2^{-j(\tau - t)}\|v\|_{\mathcal{F}_{\tau,s'}(\square)_{\mathbf{Z}}}, \quad v \in \mathcal{F}_{\tau,s'}(\square)_{\mathbf{Z}}, \tag{3.1.14}$$

for $t < \gamma,\ t \le \tau \le d$*, and*

$$\inf_{v_j \in \tilde{S}_{j,\tilde{\mathbf{Z}}}} \|v - v_j\|_{\tilde{\mathcal{F}}_{t,\tilde{s}'}(\square)_{\tilde{\mathbf{Z}}}} \lesssim 2^{-j(\tau - t)}\|v\|_{\tilde{\mathcal{F}}_{\tau,\tilde{s}'}(\square)_{\tilde{\mathbf{Z}}}}, \quad v \in \tilde{\mathcal{F}}_{\tau,\tilde{s}'}(\square)_{\tilde{\mathbf{Z}}}, \tag{3.1.15}$$

for $t \le \tilde{\gamma},\ t \le \tau \le \tilde{d}$*. Furthermore the following inverse estimates are valid*

$$\begin{aligned} \|v_j\|_{\mathcal{F}_{\tau,s'}(\square)_{\mathbf{Z}}} &\lesssim 2^{j\tau}\|v_j\|_{L_p(\square)}, \quad v_j \in S_{j,\mathbf{Z}}, \\ \|v_j\|_{\tilde{\mathcal{F}}_{\tau,\tilde{s}'}(\square)_{\tilde{\mathbf{Z}}}} &\lesssim 2^{j\tau}\|v_j\|_{L_{p'}(\square)}, \quad v_j \in \tilde{S}_{j,\tilde{\mathbf{Z}}}. \end{aligned} \tag{3.1.16}$$

(ii) *For* $\mathcal{F} = B^\tau_q(L_p)$ *as above,* $0 < \tau \le s'$ *and* $\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1$ *one has*

$$\|v\|_{\mathcal{F}(\Box)_{\mathbf{Z}}} \sim \left(\sum_{j=j_0}^{\infty} \left\{ 2^{j(\frac{n}{2} - \frac{n}{p} + \tau)} \|\langle v, \tilde{\Psi}_j^{\tilde{\mathbf{Z}}} \rangle_\Box \|_{\ell_p} \right\}^q \right)^{1/q} ,$$

$$\|v\|_{\mathcal{F}^*(\Box)_{\tilde{\mathbf{Z}}}} \sim \left(\sum_{j=j_0}^{\infty} \left\{ 2^{j(\frac{n}{2} - \frac{n}{p'} - \tau)} \|\langle v, \Psi_j^{\mathbf{Z}} \rangle_\Box \|_{\ell_{p'}} \right\}^{q'} \right)^{1/q'} . \tag{3.1.17}$$

Again analogous relations hold with interchanged roles of $\Psi^{\mathbf{Z}}, \mathbf{Z}, s'$ *and* $\tilde{\Psi}^{\tilde{\mathbf{Z}}}, \tilde{\mathbf{Z}}, \tilde{s}'$.

We will also have to make special use later of the following extreme cases of (3.1.17) which are perhaps worth being stated separately.

Remark 3.1.1. *For* $\mathcal{F}_\tau$ *as above the basis* $\Psi^{\{0,1\}^n}$ *in Properties B* (iv) *give rise to the relations*

$$\|\chi_\Box v\|_{\mathcal{F}_\tau(\mathbb{R}^n)} \sim \left(\sum_{j=j_0}^{\infty} \left\{ 2^{j(\frac{n}{2} - \frac{n}{p} + \tau)} \|\langle v, \tilde{\Psi}_j^{\emptyset} \rangle_\Box \|_{\ell_p} \right\}^q \right)^{1/q} , \quad 0 < \tau \le s' \tag{3.1.18}$$

as well as

$$\|v\|_{\tilde{\mathcal{F}}_\tau(\Box)} \sim \left(\sum_{j=j_0}^{\infty} \left\{ 2^{j(\frac{n}{2} - \frac{n}{p'} + \tau)} \|\langle v, \Psi_j^{\{0,1\}^n} \rangle_\Box \|_{\ell_{p'}} \right\}^q \right)^{1/q} , \quad 0 < \tau \le \tilde{s}'. \tag{3.1.19}$$

Remark 3.1.2. *We also remark that the validity of the first relation in* (3.1.17) *extends to the quasi-Banach spaces obtained for* $p < 1$ *which is important in the context of* nonlinear approximation [22, 26].

3.2. Wavelets on Γ_i. In order to relate the above setting on $\Box$ to the spaces $\mathcal{F}(\Gamma_i)^{\downarrow}$ we will associate with each patch Γ_i a set $\mathbf{Z}^{(i)} \subset \{0,1\}^n$ determined by the flow. In fact, the component sets $Z_\ell^{(i)}$ in $\mathbf{Z}^{(i)} = Z_1^{(i)} \times \cdots \times Z_n^{(i)}$ are characterized by

$$\kappa_i^{-1}\left(\partial^{\downarrow}\Gamma_i\right) = \bigcup_{\ell=1}^{n} [0,1]^{\ell-1} \times Z_\ell^{(i)} \times [0,1]^{n-\ell} \tag{3.2.1}$$

(where $Z_\ell^{(i)}$ could, of course, be empty); see (3.1.3). In other words, $\mathbf{Z}^{(i)}$ encodes the preimage of the inflow boundary $\partial^{\downarrow}\Gamma_i$. Accordingly, $\tilde{\mathbf{Z}}^{(i)} := \{0,1\}^n \setminus \mathbf{Z}^{(i)}$ determines the preimage of the outflow boundary $\partial^{\uparrow}\Gamma_i$ of Γ_i.

Due to the assumed regularity of the mappings κ_i the local smoothness spaces on Γ_i can be characterized by corresponding pull-backs to the parameter domain $\Box$. An immediate consequence of Remark 2.2.4 can be formulated as follows.

Remark 3.2.1. *For* $\mathbf{Z}^{(i)}$ *defined by* (3.2.1) *and any regular smooth parametrization* κ_i *of* Γ_i *one has*

$$\mathcal{F}(\Box)_{\mathbf{Z}^{(i)}} = \{v \circ \kappa_i : v \in \mathcal{F}(\Gamma_i)^{\downarrow}\} \tag{3.2.2}$$

and

$$\|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} \sim \|v \circ \kappa_i\|_{\mathcal{F}(\Box)_{\mathbf{Z}^{(i)}}}, \quad v \in \mathcal{F}(\Gamma_i)^{\downarrow}. \tag{3.2.3}$$

Completely analogous relations hold for $\mathcal{F}, \downarrow$ *replaced by* $\tilde{\mathcal{F}}, \uparrow$.

Setting $\Phi \circ \kappa_i := \{\phi \circ \kappa_i : \phi \in \Phi\}$ and defining the push-forwards onto the patches Γ_i

$$\Psi^{\Gamma_i,\downarrow} := \Psi^{\mathbf{Z}^{(i)}} \circ \kappa_i^{-1}, \quad \tilde{\Psi}^{\Gamma_i,\uparrow} := \tilde{\Psi}^{\tilde{\mathbf{Z}}^{(i)}} \circ \kappa_i^{-1}, \tag{3.2.4}$$

we immediately infer from the definition of $(\cdot,\cdot)_i$ and (3.1.7) that these collections are biorthogonal, i.e.,

$$(\Psi^{\Gamma_i,\downarrow}, \tilde{\Psi}^{\Gamma_i,\uparrow})_i = \langle \Psi^{\mathbf{Z}^{(i)}}, \tilde{\Psi}^{\tilde{\mathbf{Z}}^{(i)}} \rangle_{\Box} = \mathbf{I}. \tag{3.2.5}$$

Moreover, it follows from Remark 3.2.1 and (3.1.8) that

$$\Psi^{\Gamma_i,\downarrow} \subset \mathcal{F}_{s'}(\Gamma_i)^{\downarrow}, \quad \tilde{\Psi}^{\Gamma_i,\uparrow} \subset \tilde{\mathcal{F}}_{\tilde{s}'}(\Gamma_i)^{\uparrow}. \tag{3.2.6}$$

We can now lift the relations (3.1.17) to the patches Γ_i. Since by (2.2.3) and (3.2.4),

$$(v, \tilde{\Psi}^{\Gamma_i,\uparrow})_i = \langle v \circ \kappa_i, \tilde{\Psi}^{\tilde{\mathbf{Z}}^{(i)}} \rangle_{\Box}, \quad (v, \Psi^{\Gamma_i,\downarrow})_i = \langle v \circ \kappa_i, \Psi^{\mathbf{Z}^{(i)}} \rangle_{\Box}, \tag{3.2.7}$$

the relation (3.2.3) combined with (3.1.17) provides for $\mathcal{F} = B_q^\tau(L_p)$ (and the respective range of $\tau \le s'$)

$$\begin{aligned} \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} &\sim \left(\sum_{j=j_0}^{\infty} \left\{ 2^{j(\frac{n}{2} - \frac{n}{p} + \tau)} \|(v, \tilde{\Psi}_j^{\Gamma_i,\uparrow})_i\|_{\ell_p} \right\}^q \right)^{1/q}, \quad v \in \mathcal{F}(\Gamma_i)^{\downarrow}, \\ \|v\|_{\mathcal{F}^*(\Gamma_i)^{\uparrow}} &\sim \left(\sum_{j=j_0}^{\infty} \left\{ 2^{j(\frac{n}{2} - \frac{n}{p'} - \tau)} \|(v, \Psi_j^{\Gamma_i,\downarrow})_i\|_{\ell_{p'}} \right\}^{q'} \right)^{1/q'}, \quad v \in \mathcal{F}^*(\Gamma_i)^{\uparrow}. \end{aligned} \tag{3.2.8}$$

Here we have used that, due to the smoothness of the κ_i, the Riesz maps interrelating the inner products $(\cdot,\cdot)_i$ and $\langle \cdot,\cdot \rangle_{\Gamma_i}$ are automorphisms for all the spaces $\mathcal{F}$ under consideration.

REMARK 3.2.2. *Analogous relations hold, of course, when $\mathcal{F}(\Gamma_i)^{\downarrow}$, $\tilde{\Psi}_j^{\Gamma_i,\uparrow}$ in the first relation of* (3.2.8) *are replaced by $\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}$, $\Psi_j^{\Gamma_i,\downarrow}$, respectively. Likewise $\mathcal{F}^*(\Gamma_i)^{\uparrow}$, $\Psi_j^{\Gamma_i,\downarrow}$ in the second relation of* (3.2.8) *can be replaced by $\tilde{\mathcal{F}}^*(\Gamma_i)^{\downarrow}$, $\tilde{\Psi}_j^{\Gamma_i,\uparrow}$ and s' is replaced by $\tilde{s}'$, respectively.*

3.3. Wavelets on Γ. It is now straightforward to assemble bases on Γ. To this end, consider the inner product

$$\langle \cdot,\cdot \rangle_{\Pi} := \sum_{i=1}^{N} \langle \cdot,\cdot \rangle_{\Gamma_i} \tag{3.3.1}$$

on the product space $\prod_{i=1}^{N} L_2(\Gamma_i)$. Let T, V, S, U denote the mappings from (2.4.6) and (2.4.7). Employing in the following the abbreviation

$$\nabla^i := \nabla^{\mathbf{Z}^{(i)}}, \quad i = 1, \ldots, N,$$

for the index sets of the bases $\Psi^{\mathbf{Z}^{(i)}}, \tilde{\Psi}^{\tilde{\mathbf{Z}}^{(i)}}$, we set

$$\begin{aligned} &\psi_\lambda^\Gamma := P_i \chi_{\Gamma_i} \psi_\nu^{\Gamma_i,\downarrow}, \quad \lambda := (i,\nu),\ \nu \in \nabla^i,\ i = 1,\dots,N, \\ &\tilde{\psi}_\lambda^\Gamma := P_i^* \chi_{\Gamma_i} g_i^{-1} \tilde{\psi}_\nu^{\Gamma_i,\uparrow}, \quad \lambda := (i,\nu),\ \nu \in \nabla^i,\ i = 1,\dots,N, \end{aligned} \tag{3.3.2}$$

where g_i is defined by (2.2.2). Note that these relations can be further simplified. In fact, by (3.2.6), (2.5.13), and the definition (2.4.3) of the P_i, one has

$$\psi_\lambda^\Gamma = \chi_{\Gamma_i^\uparrow} E_i \psi_\nu^{\Gamma_i,\downarrow}, \quad \lambda = (i,\nu) \in \nabla^\Gamma. \tag{3.3.3}$$

Setting

$$\nabla^\Gamma := \bigcup_{i=1}^N (\{i\} \times \nabla^i)$$

(and analogously for the components Δ_+, ∇_-) the collections

$$\Psi^\Gamma := \{\psi_\lambda^\Gamma : \lambda \in \nabla^\Gamma\}, \quad \tilde{\Psi}^\Gamma := \{\tilde{\psi}_\lambda^\Gamma : \lambda \in \nabla^\Gamma\} \tag{3.3.4}$$

are natural candidates for wavelets on Γ.

To see this, we may identify, in view of (3.2.6), each $\psi_\nu^{\Gamma_i,\downarrow}, \nu \in \nabla^i$, with an element $\mathbf{v}_\nu^{i,\downarrow} \in \Pi\mathcal{F}^\downarrow$, obtained by setting all other components to zero. Analogously define $\mathbf{v}_\nu^{\Gamma_i,\uparrow}$ with $\psi_\nu^{\Gamma_i,\downarrow}$ replaced by $g_i^{-1}\tilde{\psi}_\nu^{\Gamma_i,\uparrow}$. By (2.4.7), this means that

$$\psi_\lambda^\Gamma = S\mathbf{v}_\nu^{i,\downarrow}, \quad \tilde{\psi}_\lambda^\Gamma = U\mathbf{v}_\nu^{i,\uparrow}, \quad \lambda = (i,\nu) \in \nabla^\Gamma; \tag{3.3.5}$$

in brief

$$\Psi^\Gamma = S\{\Psi^{\Gamma_i,\downarrow}\}_{i=1}^N, \quad \tilde{\Psi}^\Gamma = U\{g_i^{-1}\tilde{\Psi}^{\Gamma_i,\uparrow}\}_{i=1}^N. \tag{3.3.6}$$

The following observation confirms that the collections Ψ^Γ and $\tilde{\Psi}^\Gamma$ are reasonable candidates for wavelet bases on Γ.

THEOREM 3.1. *Suppose that Properties B hold for the wavelet bases on* $\Box$. *Then the collections* $\Psi^\Gamma, \tilde{\Psi}^\Gamma$ *defined by* (3.3.6) *are biorthogonal with respect to the canonical inner product on* Γ

$$\langle \Psi^\Gamma, \tilde{\Psi}^\Gamma \rangle_\Gamma = \mathbf{I}. \tag{3.3.7}$$

Proof. By (3.3.2), we obtain for $\lambda = (i,\nu), \lambda' = (l,\mu)$

$$\langle \psi_\lambda^\Gamma, \tilde{\psi}_{\lambda'}^\Gamma \rangle_\Gamma = \langle S\mathbf{v}_\nu^{i,\downarrow}, U\mathbf{v}_\mu^{l,\uparrow} \rangle_\Gamma = \langle \mathbf{v}_\nu^{i,\downarrow}, S^*U\mathbf{v}_\mu^{l,\uparrow} \rangle_\Pi.$$

In view of (2.4.17) and the fact that $P_i^* P_l^* = \delta_{i,l} P_i^*$, the right-hand side becomes

$$\langle \mathbf{v}_\nu^{i,\downarrow}, S^*U\mathbf{v}_\mu^{l,\uparrow} \rangle_\Pi = \delta_{i,l} \langle \psi_\nu^{\Gamma_i,\downarrow}, P_i^*(\chi_{\Gamma_i} g_l^{-1} \tilde{\psi}_\mu^{\Gamma_l,\uparrow}) \rangle_{\Gamma_i},$$

where we have used (2.4.7). Since, by Theorem 2.2, $(P_i^* \chi_{\Gamma_i^{\downarrow\downarrow}} v) \mid_{\Gamma_i} = v \mid_{\Gamma_i}$ for $v \mid_{\Gamma_i} \in \tilde{\mathcal{F}}(\Gamma_i)^\uparrow$, we conclude

$$\langle \psi_\lambda^\Gamma, \tilde{\psi}_{\lambda'}^\Gamma \rangle_\Gamma = \delta_{i,l} \langle \psi_\nu^{\Gamma_i,\downarrow}, g_l^{-1} \tilde{\psi}_\mu^{\Gamma_l,\uparrow} \rangle_{\Gamma_i} = \delta_{i,l} \delta_{\nu,\mu},$$

where we have first used (2.2.4) to switch from $\langle\cdot,\cdot\rangle_{\Gamma_i}$ to $(\cdot,\cdot)_i$ and then employed (3.1.7). This confirms biorthogonality (3.3.7). □

A closer look at the above construction principle (3.3.2) reveals that one can similarly realize biorthogonality relative to some modified inner product which could, for instance, be required from a particular variational formulation of an operator equation. In absence of any such specification we focus here on the standard inner product as opposed to one that depends on the construction of the wavelet bases as, e.g., in [7, 12, 21].

REMARK 3.3.1. *Recall that the wavelets on Γ_i have been defined for* any *regular parametrization of Γ_i. Their construction is in this sense completely local and* independent *of the neighboring patches. The global regularity of the manifold enters the construction of the wavelets on Γ only through the extension operators in* (3.3.3). *Only the application of these extension operators (or their adjoints) will require smoothly joining locally (re)parametrizations of neighboring patches; see sections* 4, 5 *below.*

3.4. Approximation and inverse properties, norm equivalences. We are now ready to discuss the stability properties of the bases $\Psi^\Gamma, \tilde{\Psi}^\Gamma$ on Γ. The first step is to know how accurately one can approximate by elements of the spaces

$$S_j := \operatorname{span}\{\psi_\lambda^\Gamma : \lambda \in \nabla,\ |\lambda| < j\}, \quad \tilde{S}_j := \operatorname{span}\{\tilde{\psi}_\lambda^\Gamma : \lambda \in \nabla,\ |\lambda| < j\}. \tag{3.4.1}$$

THEOREM 3.2. *Under the above assumptions one has for any $0 \le \tau \le s'$*

$$\inf_{v_j \in S_j} \|v - v_j\|_{\mathcal{F}_\tau(\Gamma)} \lesssim 2^{-j(s-\tau)} \|v\|_{\mathcal{F}_s(\Gamma)}, \quad v \in \mathcal{F}_s(\Gamma), \tag{3.4.2}$$

and similarly for $0 \le \tau \le \tilde{s}'$

$$\inf_{v_j \in \tilde{S}_j} \|v - v_j\|_{\tilde{\mathcal{F}}_\tau(\Gamma)} \lesssim 2^{-j(s-\tau)} \|v\|_{\tilde{\mathcal{F}}_s(\Gamma)}, \quad v \in \tilde{\mathcal{F}}_s(\Gamma). \tag{3.4.3}$$

Moreover, when $\mathcal{F}_{s'} = H^{s'} = B_2^{s'}(L_2)$ assume that for a given $v \in \mathcal{F}_s(\Gamma)$ the components v^i of $Tv = \{v^i\}_{i=1}^N$, respectively, of $Vv = \{v^i\}_{i=1}^N$, are smooth. Then one has

$$\inf_{v_j \in V_j} \|v - v_j\|_{L_2(\Gamma)} \lesssim 2^{-mj} \left(\sum_{i=1}^N \|v^i\|_{H^m(\Gamma)} \right) \tag{3.4.4}$$

with $m = d, \tilde{d}$, when $V_j = S_j, \tilde{S}_j$, respectively.

Proof. Recall from (3.3.2) that any $v_j = \sum_{\lambda \in \nabla^\Gamma, |\lambda| < j} d_\lambda \psi_\lambda^\Gamma$ has the form $v_j = \sum_{i=1}^N \sum_{\nu \in \nabla^i, |\nu| < j} d_{i,\nu} P_i \chi_{\Gamma_i} \psi_\nu^{\Gamma_i,\downarrow}$. Moreover, since by Theorem 2.1, (2.4.7), $v = STv = \sum_{l=1}^N P_l \chi_{\Gamma_l} v^l$ one readily infers from (2.4.8) and (2.5.13) that $(P_i v)\,|_{\Gamma_i} = v^i$. Thus Proposition 2.4.1 (i) yields

$$(P_i(v - v_j))\,|_{\Gamma_i} = v^i - \sum_{\nu \in \nabla^i, |\nu| < j} d_{i,\nu} \psi_\nu^{\Gamma_i,\downarrow}.$$

Therefore, one obtains by Theorem 2.3 for $0 \le \tau \le s' < s_\Gamma$

$$\inf_{v_j \in S_j} \|v - v_j\|_{\mathcal{F}_\tau(\Gamma)}^2 \lesssim \sum_{i=1}^N \inf_{d_{i,\nu}} \|v^i - \sum_{\nu \in \nabla^i, |\nu| < j} d_{i,\nu} \psi_\nu^{\Gamma_i,\downarrow}\|_{\mathcal{F}_\tau(\Gamma_i)^\downarrow}. \tag{3.4.5}$$

From Properties C (i), (3.1.14), and Remark 2.2.4, we infer that

$$\begin{aligned}\inf_{d_{i,\nu}} \|v^i - \sum_{\nu\in\nabla^i,|\nu|<j} d_{i,\nu}\psi_\nu^{\Gamma_i,\downarrow}\|_{\mathcal{F}_\tau(\Gamma_i)^\downarrow} &\lesssim \inf_{d_{i,\nu}} \|v^i\circ\kappa_i - \sum_{\nu\in\nabla^i,|\nu|<j} d_{i,\nu}\psi_\nu^{\mathbf{Z}^{(i)}}\|_{\mathcal{F}_\tau(\Box)_{\mathbf{Z}^{(i)}}}\\ &\lesssim 2^{-j(s-\tau)}\|v^i\circ\kappa_i\|_{\mathcal{F}_s(\Box)_{\mathbf{Z}^{(i)}}}\\ &\lesssim 2^{-j(s-\tau)}\|v^i\mid_{\Gamma_i}\|_{\mathcal{F}_s(\Gamma_i)^\downarrow}.\end{aligned}$$

Combining this with (3.4.5) and invoking Theorem 2.3 yields (3.4.2) for $s' \geq \tau \geq 0$ and $V_j = S_j$.

Let us establish now the estimates for the dual multiresolution spaces $\tilde{S}_j$. Since $g_i^{-1}\tilde{\psi}_\nu^{\Gamma_i,\uparrow} \in \tilde{\mathcal{F}}_\tau(\Gamma_i)^\uparrow$ for $\tau \leq \tilde{s}' < s_\Gamma$ the second relation of (2.5.13) in Theorem 2.2 implies that $(P_i^*\chi_{\Gamma_i}g_i^{-1}\tilde{\psi}_\nu^{\Gamma_i,\uparrow})\mid_{\Gamma_i}= g_i^{-1}\tilde{\psi}_\nu^{\Gamma_i,\uparrow}$, we conclude as above (this time using Proposition 2.4.2) that

$$\inf_{v_j\in\tilde{S}_j} \|v-\tilde{v}_j\|^2_{\mathcal{F}_\tau(\Gamma)} \lesssim \sum_{i=1}^N \inf_{d_{i,\nu}} \|(P_i^*v)\mid_{\Gamma_i} - \sum_{\nu\in\nabla^i,|\nu|<j} d_{i,\nu}g_i^{-1}\tilde{\psi}_\nu^{\Gamma_i,\uparrow}\|^2_{\tilde{\mathcal{F}}_\tau(\Gamma_i)^\uparrow}.$$

Thus it remains to estimate $\|g_i^{-1}(g_iv^i - \sum_{\nu\in\nabla^i,|\nu|<j} d_{i,\nu}\tilde{\psi}_\nu^{\Gamma_i,\uparrow})\|_{\tilde{\mathcal{F}}_\tau(\Gamma_i)^\uparrow}$ which, in view of (3.1.15), proceeds exactly as above. This confirms (3.4.3) for $s \geq \tau \geq 0$.

The rest of the assertion follows again from (3.1.14) applied to the pull-back of

$$\inf_{d_{i,\nu}} \|v^i - \sum_{\nu\in\nabla^i,|\nu|<j} d_{i,\nu}\psi_\nu^{\Gamma_i,\downarrow}\|_{\mathcal{F}_\tau(\Gamma_i)^\downarrow} \lesssim 2^{-j(d-\tau)}\|v^i\|_{\mathcal{F}_{d,s'}(\Gamma_i)},$$

where $\mathcal{F}_{d,s'}(\Gamma_i)$ is the push-forward of $\mathcal{F}_{d,s'}(\Box)_{\mathbf{Z}^{(i)}} = H^d(\Box)\cap\mathcal{F}_{s'}(\Box)_{\mathbf{Z}^{(i)}}$. □

Remark 3.4.1. *When Γ had a boundary, one could have incorporated homogeneous Dirichlet boundary conditions on all or part of the boundary composed of a union of patch boundaries. It is obvious from the above proof that analogous estimates persist to hold as long as the approximands satisfy corresponding boundary conditions. Details are left to the reader.*

There is the following counterpart to estimates of the type (3.4.2).

Remark 3.4.2. *One has*

$$\|v_j\|_{\mathcal{F}_{\tau,s'}(\Gamma)} \lesssim 2^{\tau j}\|v_j\|_{L_p(\Gamma)}, \quad v_j \in S_j,\ 0\leq\tau<\min\{s_\Gamma,\gamma\}. \tag{3.4.6}$$

An analogous estimate holds for the spaces $\tilde{S}_j$ with $\mathcal{F}, p, s', \gamma$, replaced by $\tilde{\mathcal{F}}, p', \tilde{s}', \tilde{\gamma}$.

Proof. Properties C (i) (3.1.16) ensures that

$$\|v_j\|_{\mathcal{F}_{\tau,s'}(\Box)_{\mathbf{Z}^{(i)}}} \lesssim 2^{\tau j}\|v_j\|_{L_p(\Box)}, \quad v_j \in S_{j,\mathbf{Z}^{(i)}}\ \forall\ 0\leq\tau\leq\gamma. \tag{3.4.7}$$

Analogous estimates for the spaces $\mathcal{F}_\tau(\Gamma_i)^\downarrow$ follow immediately from Remark 3.2.1. Now one can invoke Theorem 2.3 to confirm (3.4.6). □

Because of its importance we consider first the special case $\mathcal{F}_\tau = H^\tau$, i.e., $p = q = 2$. Combining Theorem 3.2 with Remark 3.4.2 we can apply Corollary 5.2 in [14] to conclude the following Sobolev norm equivalences for a *whole regularity range* which play a key role in wavelet concepts for the numerical solution of operator equations; see section 5.

THEOREM 3.3. *Under the above assumptions one has for any $v \in H^\tau(\Gamma)$*

$$\|v\|_{H^\tau(\Gamma)} \sim \left(\sum_{\lambda \in \nabla^\Gamma} 2^{2\tau|\lambda|} |\langle v, \tilde{\psi}_\lambda^\Gamma \rangle_\Gamma|^2 \right)^{1/2}, \quad \tau \in [-\tilde{s}', s']; \tag{3.4.8}$$

see (3.1.12), *where for $\tau < 0$ we mean as usual $H^\tau(\Gamma) = (H^{-\tau}(\Gamma))^*$. An analogous relation holds with $\tilde{\psi}_\lambda^\Gamma$ replaced by ψ_λ^Γ and reversed end points of the interval describing the range of validity. In particular, $\Psi^\Gamma, \tilde{\Psi}^\Gamma$ are Riesz bases for $L_2(\Gamma)$.*

Again we could have formulated analogous relations for Sobolev spaces $H^s_{0,D}(\Gamma)$ satisfying homogeneous Dirichlet boundary conditions on some part D of the boundary of Γ.

REMARK 3.4.3. *In the case $p = 2 = q$, i.e., $\mathcal{F}_s = H^s$ it is important to estimate the error in dual norms; see* [19]. *In particular, we have for $-\sigma \leq \tau \leq \sigma$*

$$\inf_{v_j \in V_j} \|v - v_j\|_{H^\tau(\Gamma)} \lesssim 2^{-j(s-\tau)} \|v\|_{H^s(\Gamma)}, \quad v \in H^s(\Gamma), \tag{3.4.9}$$

where $\sigma = s' < \gamma, \tilde{s}' < \tilde{\gamma}$ for $V_j = S_j, \tilde{S}_j$, respectively.

Proof. As usual one can use duality for the case $\tau < 0$. In fact, let

$$Q_j v := \sum_{\lambda \in \nabla^\Gamma} \langle v, \tilde{\psi}_\lambda^\Gamma \rangle_\Gamma \psi_\lambda^\Gamma, \quad Q_j^* v := \sum_{\lambda \in \nabla^\Gamma} \langle v, \psi_\lambda^\Gamma \rangle_\Gamma \tilde{\psi}_\lambda^\Gamma,$$

and observe that, due to the uniform boundedness of the Q_j, Q_j^* in $H^\tau(\Gamma)$, $0 \leq \tau \leq s', \tilde{s}'$, which follows from Theorem 3.3,

$$\|v - Q_j v\|_{H^\tau(\Gamma)} \lesssim \inf_{v_j \in S_j} \|v - v_j\|_{H^\tau(\Gamma)}, \quad \|v - Q_j^* v\|_{H^\tau(\Gamma)} \lesssim \inf_{\tilde{v}_j \in \tilde{S}_j} \|v - \tilde{v}_j\|_{H^\tau(\Gamma)}.$$

Now suppose that $0 < \tau \leq s$ and note that

$$\begin{aligned} \|v - Q_j v\|_{H^{-\tau}(\Gamma)} &= \sup_{\|w\|_{H^\tau(\Gamma)}=1} \langle Q_j v - v, w \rangle_\Gamma = \sup_{\|w\|_{H^\tau(\Gamma)}=1} \langle (Q_j - I)^2 v, w \rangle_\Gamma \\ &= \sup_{\|w\|_{H^\tau(\Gamma)}=1} \langle Q_j v - v, Q_j^* w - w \rangle_\Gamma \\ &\leq \sup_{\|w\|_{H^\tau(\Gamma)}=1} \|Q_j v - v\|_{L_2(\Gamma)} \|w - Q_j^* w\|_{L_2(\Gamma)} \\ &\lesssim \sup_{\|w\|_{H^\tau(\Gamma)}=1} \left(2^{-j\tau} \|w\|_{H^\tau(\Gamma)} 2^{-js} \|v\|_{H^s(\Gamma)} \right), \end{aligned}$$

where we have used the previously established estimates for positive Sobolev indices. The reasoning for $\tilde{S}_j$ is completely analogous. □

We conclude this section with similar norm equivalences for the general case of Besov spaces $\mathcal{F}_\tau = B^\tau_q(L_p)$. This time we combine the local relations (3.2.8) with Theorem 2.3 to derive similar relations for Γ.

THEOREM 3.4. *One has for $\mathcal{F} = B^\tau_q(L_p)$ and $\frac{1}{p} + \frac{1}{p'} = \frac{1}{q} + \frac{1}{q'} = 1$, $0 < \tau \leq s'$,*

$$\|v\|_{\mathcal{F}(\Gamma)} \sim \left(\sum_{j=j_0}^\infty \left\{ 2^{j(\frac{n}{2} - \frac{n}{p} + \tau)} \|\langle v, \tilde{\Psi}_j^\Gamma \rangle_\Gamma\|_{\ell_p} \right\}^q \right)^{1/q}, \quad v \in \mathcal{F}(\Gamma), \tag{3.4.10}$$

and

$$\text{(3.4.11)}\quad \|v\|_{\mathcal{F}^*(\Gamma)} \sim \left(\sum_{j=j_0}^{\infty}\left\{2^{j(\frac{d}{2}-\frac{d}{p'}-\tau)}\|\langle v,\Psi_j^{\Gamma}\rangle_{\Gamma}\|_{\ell_{p'}}\right\}^{q'}\right)^{1/q'}, \qquad v\in\mathcal{F}^*(\Gamma).$$

Analogous relations hold when interchanging the roles of Ψ^{Γ}, s' *and* $\tilde{\Psi}^{\Gamma}, \tilde{s}'$.

Proof. By Theorem 2.3, we have

$$\text{(3.4.12)}\qquad \|v\|_{\mathcal{F}(\Gamma)} \sim \left(\sum_{i=1}^{N}\|(P_i v)\,|_{\Gamma_i}\,\|^q_{\mathcal{F}(\Gamma_i)^{\downarrow}}\right)^{1/q}.$$

Invoking (3.2.8), one obtains

$$\text{(3.4.13)}\qquad \|(P_i v)\,|_{\Gamma_i}\,\|^q_{\mathcal{F}(\Gamma_i)^{\downarrow}} \sim \sum_{j=j_0}^{\infty}\left\{2^{j(\frac{n}{2}-\frac{n}{p}+\tau)}\|(P_i v,\tilde{\Psi}_j^{\Gamma_i,\uparrow})_i\|_{\ell_p}\right\}^q.$$

Since again, by (2.2.4) and (3.3.2), analogous reasoning to the above yields

$$\text{(3.4.14)}\qquad \begin{aligned}(P_i v,\tilde{\psi}_{\nu}^{\Gamma_i,\uparrow})_i &= \langle P_i v, g_i^{-1}\tilde{\psi}_{\nu}^{\Gamma_i,\uparrow}\rangle_{\Gamma_i} = \langle P_i v, \chi_{\Gamma_i} g_i^{-1}\tilde{\psi}_{\nu}^{\Gamma_i,\uparrow}\rangle_{\Gamma}\\ &= \langle v, P_i^*\chi_{\Gamma_i} g_i^{-1}\tilde{\psi}_{\nu}^{\Gamma_i,\uparrow}\rangle_{\Gamma} = \langle v,\tilde{\psi}^{\Gamma}_{(i,\nu)}\rangle_{\Gamma},\end{aligned}$$

and combining (3.4.12), (3.4.13), and (3.4.14) provides (3.4.10). The equivalence relation (3.4.11) follows by analogous arguments based on (2.5.19) in Theorem 2.3 and again in (3.2.8). □

Note that we could have recovered Theorem 3.3 from Theorem 3.4 combined with an interpolation and duality argument; see [14]. Accordingly, we would get for $\tau = 0$ a norm equivalence for $B^0_q(L_p(\Gamma))$ which differs from L_p when $p \neq 2$.

REMARK 3.4.4. *Again if one dispenses with duality relations, the validity of equivalences of the type* (3.4.10) *(for positive regularity index τ) extends to $p < 1$ which follows from Remark* 3.1.2 *and a corresponding version of the first relation in* (2.5.17) *of Theorem* 2.3.

4. Extensions. The above construction and, in particular, its *practicality* hinges on the identification of suitable extension operators. This section is devoted to this issue. We will depart from the development in [9] by interrelating the construction of extension operators directly with the wavelet bases on □. This will result in *scale-dependent* extensions. First we need a few preparations.

4.1. Lifted extensions. In view of the preceding development it is natural to construct extensions in the parameter domain and then lift them to the manifold. To this end, we adhere to the notation in section 3.1 and recall the meaning of $\square_{\tilde{\mathbf{Z}}^{(i)}}$ from (3.2.1) and (3.1.2). At this point we have to interrelate the parametrizations of the patches Γ_i with the global smoothness of the manifold. For every patch Γ_i this will concern only κ_i and the parametrizations of *neighboring* patches. In particular, at *no point* will our approach ever require any specific *global* parametric representation of Γ. We will address this issue in two steps. First we identify the conceptually relevant property which holds in full generality, then we formulate a more specified version which is desirable from a practical point of view. Since the validity of this latter property has been verified for various concrete constructions only for $n \leq 2$, we prefer to state it separately.

Property G. *There exists a neighborhood $\diamond \subset \mathbb{R}^n$ of $\square$, i.e.,* $\operatorname{dist}(\partial\square, \partial\diamond) \geq a > 0$, *such that for each $i = 1, \dots, N$ one can find a neighborhood $\hat{\Gamma}_i$ of Γ_i in the above sense and a parametric mapping $\hat{\kappa}_i$ such that*

$$\hat{\kappa}_i(\diamond) = \hat{\Gamma}_i, \quad \hat{\kappa}_i, \hat{\kappa}_i^{-1} \in C^{m,1}, \quad \hat{\kappa}_i|_{\square} = \kappa_i. \tag{4.1.1}$$

In brief, each κ_i can be extended to a $C^{m,1}$-homeomorphism to some open neighborhood of Γ_i.

We emphasize first that Property G entails *no* constraints on the topology of Γ. In fact, recall that our smoothness assumption on Γ means that there exists a covering $\mathcal{C}$ of Γ of neighborhoods Γ' which are $C^{m,1}$-homeomorphic images of neigborhoods in $\mathbb{R}^n$. Since Γ is compact, one can subdivide the patch complex $\{\Gamma_i\}$ finiteley many times, e.g., by dyadic subdivisions of the parameter domain $\square$, such that the closure of each of the resulting cubical patches on Γ is fully contained in some neighborhood in $\mathcal{C}$. Thus, independently of the topology of Γ, there exists a $C^{m,1}$ parametrization of a neighborhood of each patch in the (subdivided) complex as stated in Property G.

Of course, this argument is not very practical. A little more effort shows that extensibility in the sense of (4.1.1) can be confirmed without subdividing the patch complex. A constructive argument can be based on building first a $C^{m,1}$-extension of a given κ_i which takes a neigborhood of $\square$ into a $C^{m,1}$ parametric extension of Γ_i staying close to Γ. A suitable "smooth correction" yields then a $C^{m,1}$ parametric representation of a neighborhood of Γ_i in Γ. While Property G is important for theoretical reasons it is not quite satisfactory from a practical point of view so that we dispense here with going into more details of the argument.

A practical counterpart to Property G that in the framework of [31, 42, 44] is realized for *arbitrary topology* can now be formulated as follows.

Property G′. *For any $i \in \{1, \dots, N\}$, let*

$$\hat{\Gamma}_i := \bigcup \{\overline{\Gamma}_l : \overline{\Gamma}_l \cap \overline{\Gamma}_i \neq \emptyset\}$$

denote the union of all patches whose closure intersects the closure of Γ_i. If $\hat{\Gamma}_i$ is topologically equivalent to a domain in $\mathbb{R}^n$, then there exists a neighborhood $\diamond \subset \mathbb{R}^n$ of $\square$ and a parametric mapping $\hat{\kappa}_i$ such that

$$\hat{\kappa}_i(\diamond) = \hat{\Gamma}_i, \quad \hat{\kappa}_i \hat{\kappa}_i^{-1} \in C^{m,1}, \quad \hat{\kappa}_i|_{\Gamma_i} = \kappa_i. \tag{4.1.2}$$

Moreover, $\diamond$ is a disjoint union of subdomains of the form $\rho_l(\square)$, ($\rho_i(\square) = \square$) such that

$$\hat{\kappa}_i(\rho_l(\square)) = \Gamma_l, \tag{4.1.3}$$

i.e., the restriction of $\hat{\kappa}_i$ to $\rho_l(\square)$ is a regular reparametrization of κ_l.

Of course, the assumption that $\hat{\Gamma}_i$ is topologically equivalent to a domain in $\mathbb{R}^n$ is no serious restriction and can always be enforced by one subdivision step (which may be needed for separating singular vertices anyway, recall the discussion in section 2.1). We will therefore assume throughout the following that this holds for each $\hat{\Gamma}_i$.

For simplicity, assume now the validity of Property G′ in the general case. First recall that $\square_{\tilde{\mathbf{Z}}^{(i)}}$, $\Gamma_i^{\uparrow}$ are subsets of $\diamond$ and $\hat{\Gamma}_i$, respectively. Therefore let $\kappa_i^{\uparrow}$ denote the restriction of $\hat{\kappa}_i$ to $\square_{\tilde{\mathbf{Z}}^{(i)}}$, i.e.,

$$\kappa_i^{\uparrow}(\square_{\tilde{\mathbf{Z}}^{(i)}}) = \Gamma_i^{\uparrow}, \quad \kappa_i^{\uparrow}\Big|_{\square} = \kappa_i. \tag{4.1.4}$$

We could also work with Property G in which case $\Gamma_i^\uparrow$ has to be replaced by a suitable subset that still contains the relative interior of the outflow boundary (employing reparametrizations one does not have to restrict the domain $\square_{\tilde{\mathbf{Z}}^{(i)}}$). Of course, the smoothness of $\kappa_i^\uparrow$ (and analogously of $\kappa_i^\downarrow$) is now limited by the global smoothness of Γ.

Note that each mapping $\kappa_i : \square \to \Gamma_i$ (and likewise $\kappa_i^\uparrow$) induces a mapping $\kappa_i^* : \mathcal{F}(\Gamma_i) \to \mathcal{F}(\square)$ by

$$(\kappa_i^* v)(y) = v(\kappa_i(y)), \quad y \in \square. \tag{4.1.5}$$

Now suppose that $A_{\tilde{\mathbf{Z}}^{(i)}}$ is an extension from $\mathcal{F}(\square)$ into $\mathcal{F}(\square_{\tilde{\mathbf{Z}}^{(i)}})$ satisfying (2.5.7) and (2.5.8) and define

$$E_i = (\kappa_i^\uparrow)^{*\,-1} A_{\tilde{\mathbf{Z}}^{(i)}} \kappa_i^*, \quad i = 1, \ldots, N. \tag{4.1.6}$$

It is not hard to verify that then the E_i also satisfy (2.5.7) and (2.5.8).

Straightforward calculations show that when E_i is defined by (4.1.6) then the adjoints E_i^* are given by

$$E_i^* = (\kappa_i^{-1})^* \left(|\partial \kappa_i|^{-1} A^*_{\tilde{\mathbf{Z}}^{(i)}} (|\partial \kappa_i^\uparrow| (\kappa_i^\uparrow)^* \right). \tag{4.1.7}$$

It will be instructive to consider now the following concrete type of extensions used in [9] which will serve as *one possible* building block.

4.2. Hestenes extensions. Since all the domains appearing in the above construction are cubes, hyperrectangles, and their parametric images, it is natural to employ tensor products of extension operators for the unit interval. Following [9], suitable versions of Hestenes extensions appear to be a possible choice which will be pointed out first since it may be used as a starting point for subsequent modifications to be explained in more detail later. To this end, choose for some $l \in \mathbb{N}$ real numbers β_i such that

$$-2 \le \beta_1 < \cdots < \beta_l \le -\frac{1}{2}$$

and assume that $\eta \in C^\infty(\mathbb{R})$ satisfies

$$\eta(x) = \begin{cases} 1, & x \in \left[-\frac{1}{4}, \frac{1}{4}\right], \\ 0, & x \notin (-\frac{1}{2}, \frac{1}{2}). \end{cases} \tag{4.2.1}$$

Of course, the support of the cut-off function η could be chosen to be much smaller. The present choice refers for simplicity to Property G′ which is relevant for practical applications. Adaptations to Property G are straightforward and left to the reader. Now define

$$\begin{aligned} (A_Z f)(x) := {}& \chi_{[0,1]}(x) f(x) + \chi_{[-1,0] \cap [0,1]_Z}(x) \sum_{j=1}^{l} \alpha_j (\eta f)(\beta_j x) \\ & + \chi_{[1,2] \cap [0,1]_Z}(x) \sum_{j=1}^{l} \alpha_j \eta(\beta_j(x-1)) f(1 + \beta_j(x-1)), \end{aligned} \tag{4.2.2}$$

where the numbers $\alpha_j, j = 1, \dots l$, satisfy for $l \geq 2m+2$

$$\sum_{j=1}^{l} \alpha_j \beta_j^k = 1, \quad k = -m-1, \dots, m. \tag{4.2.3}$$

Hence, whenever $0 \in \tilde{Z}$, one has

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^k (A_Z f)(x)\Big|_{x=0^-} = \sum_{j=1}^{l} \alpha_j \beta_j^k f^{(k)}(0) = f^{(k)}(0), \quad k = 0, \dots, m, \tag{4.2.4}$$

and, when $1 \in \tilde{Z}$,

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^k (A_Z f)(x)\Big|_{x=1^+} = \sum_{j=1}^{l} \alpha_j \beta_j^k f^{(k)}(1) = f^{(k)}(1), \quad k = 0, \dots, m. \tag{4.2.5}$$

Here we have also used that for $2 \geq x \geq 1$ one has $\beta_j x \leq -\frac{1}{2}$, $j = 1, \dots, l$, so that, by (4.2.1), $\eta(\beta_j x) = 0$, $x \geq 1$. Likewise, when $x \leq 0$ one has $\beta_j(x-1) \geq \frac{1}{2}$, $j = 1, \dots, l$, so that $\eta(\beta_j(x-1)) = 0$ for $x \leq 0$.

Next note that

$$\begin{aligned}
\int\limits_{\mathbb{R}} (A_Z f)(x) g(x)\,\mathrm{d}x &= \int\limits_{[0,1]_Z} (A_Z f)(x) g(x)\,\mathrm{d}x \\
&= \int\limits_0^1 f(x) g(x)\,\mathrm{d}x + \int\limits_{[-1,0]\cap[0,1]_Z} \sum_{j=1}^{l} \alpha_j \eta(\beta_j x) f(\beta_j x) g(x)\,\mathrm{d}x \\
&\quad + \int\limits_{[1,2]\cap[0,1]_Z} \sum_{j=1}^{l} \alpha_j \eta(\beta_j(x-1)) f(1+\beta_j(x-1)) g(x)\,\mathrm{d}x.
\end{aligned}$$

Straightforward computation reveals that then

$$\begin{aligned}
(A_Z^* g)(x) &= \chi_{[0,1]}(x) g(x) \\
&\qquad -\chi_{[-1,0]\cap[0,1]_Z}(x-1) \sum_{j=1}^{l} \beta_j^{-1} \alpha_j \eta(x) g(\beta_j^{-1} x) \\
&\qquad -\chi_{[1,2]\cap[0,1]_Z}(1+x) \sum_{j=1}^{l} \beta_j^{-1} \alpha_j \eta(x-1) g(1+\beta_j^{-1}(x-1)).
\end{aligned} \tag{4.2.6}$$

Here we have used the choice of the support of the cut-off function η.

As before, the choice of the α_j, $j = 1, \dots, l$, ensures that when $0 \in Z$, resp., $1 \in Z$,

$$\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^\ell (A_Z^* g)(0) = 0, \quad \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^\ell (A_Z^* g)(1) = 0, \quad \ell = 0, \dots, m. \tag{4.2.7}$$

The next step is to form tensor products of these operators. To this end, let for $Z_\ell \subseteq \{0,1\}$, $\ell = 1, \dots, n$,

$$\mathbf{Z} := Z_1 \times \cdots \times Z_n$$

and define

$$A_{\mathbf{Z}} f = (A_{Z_1} \otimes \cdots \otimes A_{Z_n})\, f. \tag{4.2.8}$$

Here we set

$$(A_{Z_i} f)(x_1, \ldots, x_n) = (A_{Z_i} f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n))(x_i)$$

and for $i \le n$

$$\begin{aligned} &((A_{Z_i} \otimes \cdots \otimes A_{Z_1})\, f)\, (x_1, \ldots, x_n) \\ &\quad = \big(A_{Z_i} \left((A_{Z_{i-1}} \otimes \cdots \otimes A_{Z_1})\, f\right) (x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n)\big)\, (x_i). \end{aligned} \tag{4.2.9}$$

One can verify that for any permutation π of $\{1, \ldots, n\}$

$$A_{Z_1} \otimes \cdots \otimes A_{Z_n} = A_{Z_{\pi(1)}} \otimes \cdots \otimes A_{Z_{\pi(n)}}. \tag{4.2.10}$$

It remains to *lift* these extension operators to the manifold Γ as described in section 4.1. Recall from Property G that when Γ is a $C^{m,1}$-surface κ_i then $\kappa_i^{\uparrow}$ and its inverse can be chosen to be $C^{m,1}$, so that by construction, $\hat{E}_i$ defined by (4.1.6) in combination with the Hestenes extension (4.2.8) satisfies

$$\hat{E}_i v \in C^{m,1}(\Gamma_i^{\uparrow}) \text{ if } v \in C^{m,1}(\Gamma_i). \tag{4.2.11}$$

Note that for applications in connection with boundary integral equations the case $m = 0$ is of particular interest so that $l = 2$ in (4.2.2) suffices.

4.3. Scale-dependent modifications. From a practical point of view the above Hestenes extensions still have certain drawbacks. To explain this suppose that f is smooth and has support strictly contained in $(0, 1/4)$, say. Thus the trivial extension of f to $\mathbb{R}$ by zero would satisfy (4.2.5). However, it is clear that A_Z defined by (4.2.2) may very well differ from zero outside $(0, 1)$. Therefore, even though by Proposition 2.4.1 (ii) and (2.5.14) in Theorem 2.2,

$$\chi_{\Gamma_i} (P_i v) \,|_{\Gamma_i} = \chi_{\Gamma_i} v \in \mathcal{F}(\Gamma) \quad \text{for} \;\; v \in \mathcal{F}(\Gamma_i)^{\downarrow} \cap \mathcal{F}(\Gamma_i)^{\uparrow}, \tag{4.3.1}$$

the above property of the extensions E_i may cause that

$$P_i v \,|_{\Gamma \setminus \Gamma_i} \not\equiv 0, \tag{4.3.2}$$

i.e., $P_i v$ may differ from $\chi_{\Gamma_i} v$ on $\Gamma_i^{\uparrow}$ and hence on all of Γ.

In this section we will point out how to construct modified extension operators which do not suffer from this deficiency. To this end, we will assume throughout this section that $\hat{E}_i$ are *some* (initial) extension operators satisfying Assumption A (2.5.7), (2.5.8) (for instance, Hestenes extensions of the type discussed in section 4.2).

The next result says that one can always construct extensions also satisfying Assumption A in such a way that those wavelets which do not interfere with the outflow boundary are extended outside Γ_i by zero. To this end, recall from Properties B (iv) the sets

$$\Lambda_i^{\uparrow} := \Lambda_{\mathbf{Z}^{(i)}}^{\{0,1\}^n}, \tag{4.3.3}$$

where as before $\mathbf{Z}^{(i)}$ is related to Γ_i by (3.2.1).

Theorem 4.1. *Let $\mathcal{F} = \mathcal{F}_{s'}$. For $\hat{E}_i$ satisfying* (2.5.7), (2.5.8) *let*

$$E_i v := \sum_{\nu \in \nabla^i \setminus \Lambda_i^{\uparrow}} (v, \tilde{\psi}_\nu^{\Gamma_i,\uparrow})_i \hat{E}_i \psi_\nu^{\Gamma_i,\downarrow} + \sum_{\nu \in \Lambda_i^{\uparrow}} (v, \tilde{\psi}_\nu^{\Gamma_i,\uparrow})_i \chi_{\Gamma_i} \psi_\nu^{\Gamma_i,\downarrow}. \tag{4.3.4}$$

Then one has

$$\|E_i v\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} \lesssim \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}}, \quad v \in \mathcal{F}(\Gamma_i)^{\downarrow}. \tag{4.3.5}$$

Moreover, the corresponding adjoints E_i^* satisfy (2.5.8) *now for $\tau \le \tilde{s}'$, i.e., the new extensions E_i also satisfy Assumption A.*

Proof. One immediately infers from (3.2.6) and the definition of $\Lambda_i^{\uparrow}$ that for any $v \in \mathcal{F}(\Gamma_i)^{\downarrow}$

$$v_0 := \sum_{\nu \in \Lambda_i^{\uparrow}} (v, \tilde{\psi}_\nu^{\Gamma_i,\uparrow})_i \chi_{\Gamma_i} \psi_\nu^{\Gamma_i,\downarrow} \in \mathcal{F}(\Gamma_i)^{\downarrow} \cap \mathcal{F}(\Gamma_i)^{\uparrow}.$$

By construction, one has $E_i v_0 = \chi_{\Gamma_i} v_0$ and

$$E_i(v - v_0) = \hat{E}_i(v - v_0), \tag{4.3.6}$$

so that we have

$$\|E_i v_0\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} = \|v_0\|_{\mathcal{F}(\Gamma_i)^{\uparrow\downarrow}}.$$

By Properties B (iv) and the definition of $\Lambda_i^{\uparrow}$, we can employ Remark 3.1.1 and (3.1.18) to estimate $\|v_0\|_{\mathcal{F}(\Gamma_i)^{\uparrow\downarrow}}$ by the discrete norm relative to $\Psi^{\{0,1\}^n}$. Since the summands appearing in this expression are a subset of the terms occurring in the discrete norm (3.2.8) of v expanded with respect to $\Psi^{\Gamma_i,\downarrow}$, we have

$$\|v_0\|_{\mathcal{F}(\Gamma_i)^{\uparrow\downarrow}} \lesssim \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}}. \tag{4.3.7}$$

Furthermore, by Assumptions A (2.5.7) and (4.3.6),

$$\|E_i(v - v_0)\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} = \|\hat{E}_i(v - v_0)\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} \lesssim \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} + \|v_0\|_{\mathcal{F}(\Gamma_i)^{\downarrow}}.$$

Since $\|v_0\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} \lesssim \|v_0\|_{\mathcal{F}(\Gamma_i)^{\downarrow,\uparrow}}$ we can invoke (4.3.7) again to conclude finally by (4.3.4),

$$\begin{aligned} \|E_i v\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} &\le \|E_i(v - v_0)\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} + \|E_i v_0\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} \\ &\lesssim \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} + 2\|v_0\|_{\mathcal{F}(\Gamma_i)^{\downarrow}} \lesssim \|v\|_{\mathcal{F}(\Gamma_i)^{\downarrow}}, \end{aligned}$$

which confirms (4.3.5).

As for the rest of the claim, note that for $w \in \tilde{\mathcal{F}}_{\tilde{s}'}(\Gamma_i^{\uparrow})$

$$\begin{aligned} \langle E_i v, w \rangle_{\Gamma_i^{\uparrow}} &= \sum_{\nu \in \nabla^i \setminus \Lambda_i^{\uparrow}} (v, \tilde{\psi}_\nu^{\Gamma_i,\uparrow})_i \langle \hat{E}_i \psi_\nu^{\Gamma_i,\downarrow}, w \rangle_{\Gamma_i^{\uparrow}} + \sum_{\nu \in \Lambda_i^{\uparrow}} (v, \tilde{\psi}_\nu^{\Gamma_i,\uparrow})_i \langle \psi_\nu^{\Gamma_i,\downarrow}, w \rangle_{\Gamma_i} \\ &= \left(v, \sum_{\nu \in \nabla^i \setminus \Lambda_i^{\uparrow}} \langle \hat{E}_i^* w, \psi_\nu^{\Gamma_i,\downarrow} \rangle_{\Gamma_i} \tilde{\psi}_\nu^{\Gamma_i,\uparrow} + \sum_{\nu \in \Lambda_i^{\uparrow}} \langle w, \psi_\nu^{\Gamma_i,\downarrow} \rangle_{\Gamma_i} \tilde{\psi}_\nu^{\Gamma_i,\uparrow} \right)_i \\ &= \left\langle v, g_i^{-1} \left\{ \sum_{\nu \in \nabla^i \setminus \Lambda_i^{\uparrow}} (g_i \hat{E}_i^* w, \psi_\nu^{\Gamma_i,\downarrow})_i \tilde{\psi}_\nu^{\Gamma_i,\uparrow} + \sum_{\nu \in \Lambda_i^{\uparrow}} (g_i w, \psi_\nu^{\Gamma_i,\downarrow})_i \tilde{\psi}_\nu^{\Gamma_i,\uparrow} \right\} \right\rangle_{\Gamma_i}, \end{aligned}$$

i.e.,

$$E_i^* w \;=\; g_i^{-1}\left\{\sum_{\nu\in\nabla^i\setminus\Lambda_i^{\uparrow}} (g_i\hat{E}_i^* w, \psi_\nu^{\Gamma_i,\downarrow})_i\tilde{\psi}_\nu^{\Gamma_i,\uparrow} + \sum_{\nu\in\Lambda_i^{\uparrow}} (g_i w, \psi_\nu^{\Gamma_i,\downarrow})_i\tilde{\psi}_\nu^{\Gamma_i,\uparrow}\right\}$$

$$=: w_1 + w_0. \tag{4.3.8}$$

By assumption (2.5.8) on $\hat{E}_i$, one has $\hat{E}_i^* w \in \tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}$ since $\tilde{s}' \leq s$; see (3.1.12). By the obvious analog to (3.2.8) and Assumption A (2.5.8), one has

$$\|g_i w_1\|_{\tilde{\mathcal{F}}_{\tilde{s}'}(\Gamma_i)^{\uparrow}} \;\lesssim\; \|g_i\hat{E}_i^* w\|_{\tilde{\mathcal{F}}_{\tilde{s}'}(\Gamma_i)^{\uparrow}} \;\lesssim\; \|w\|_{\tilde{\mathcal{F}}(\Gamma_i^{\uparrow})}. \tag{4.3.9}$$

By Remark 3.2.2 and the definition of $\Lambda_i^{\uparrow}$ in Properties B (iv), $\|w_0\|_{\tilde{\mathcal{F}}(\Gamma_i)^{\uparrow}}$ can be estimated by a discrete norm analogous to the first relation in (3.2.8). This discrete norm involves a subset of coefficients of w expanded with respect to the basis $\tilde{\Psi}^{\emptyset}$. Thus by (3.1.19) in Remark 3.1.1, this, in turn, can be estimated by $\|w\|_{\tilde{\mathcal{F}}(\Gamma_i)} \leq \|w\|_{\tilde{\mathcal{F}}(\Gamma_i^{\uparrow})}$ which together with (4.3.9) confirms (2.5.8) and finishes the proof. □

Corollary 4.3.1. *Let P_i be defined by* (2.4.2) *and* (2.4.3) *with respect to the E_i defined by* (4.3.4)*. Then*

$$P_i\left(\chi_{\Gamma_i}\psi_\nu^{\Gamma_i,\downarrow}\right) = \chi_{\Gamma_i}\psi_\nu^{\Gamma_i,\downarrow},\quad \nu\in\Lambda_i^{\uparrow},\quad P_i\left(\psi_\nu^{\Gamma_i,\downarrow}\right) = \hat{E}_i\psi_\nu^{\Gamma_i,\downarrow},\quad \nu\in\nabla^i\setminus\Lambda_i^{\uparrow}. \tag{4.3.10}$$

Proof. The relations (4.3.10) follow from Theorem 2.2 (see (2.5.13)), (2.4.3), and the definition (4.3.4). □

We conclude this section by quantifying the locality of the extensions E_i expressed in (2.5.6). We will require throughout the paper that

$$\operatorname{diam}(\operatorname{supp} E_i v) \;\lesssim\; \operatorname{diam}(\operatorname{supp} v) \tag{4.3.11}$$

holds. This is obviously the case for the Hestenes extensions as well as for the scale-dependent extensions based on *any* extension satisfying (4.3.11). As a consequence, we have, in view of (3.1.10), that

$$\operatorname{diam}(\operatorname{supp}\psi_\lambda^{\Gamma}),\ \operatorname{diam}(\operatorname{supp}\tilde{\psi}_\lambda^{\Gamma}) \;\lesssim\; 2^{-|\lambda|}. \tag{4.3.12}$$

5. Discretization of operators.

This section is devoted to a brief outline of some practical consequences of the foregoing constructions. More details will be given in [23] in the context of a more specified class of boundary integral operators.

5.1. A global point of view.

Suppose throughout the rest of the paper that $t \in \mathbb{R}$ satisfies $|t| \leq s < s_\Gamma$ so that $\mathcal{F}(\Gamma) = H^t(\Gamma)$ is well defined. When $t < 0$ the space $H^t(\Gamma)$ is to be understood as $(H^{-t}(\Gamma))^*$ (which agrees with the conventional notation when Γ has no boundary). Let $\mathcal{L} : H^t(\Gamma) \to H^{-t}(\Gamma)$ be a linear boundedly invertible operator, i.e.,

$$\|\mathcal{L}v\|_{H^{-t}(\Gamma)} \sim \|v\|_{H^t(\Gamma)},\quad v \in H^t(\Gamma). \tag{5.1.1}$$

Therefore for any $f \in H^{-t}(\Gamma)$ the equation

$$\mathcal{L}u = f \tag{5.1.2}$$

has a unique solution $u \in H^t(\Gamma)$. This covers a large class of (not necessarily symmetric) *elliptic* integral as well as differential equations (see, e.g., [16, 45]).

We wish to study Galerkin discretizations of (5.1.2) involving the above bases $\Psi^\Gamma, \tilde{\Psi}^\Gamma$. To describe this, let us denote for $\Lambda \subset \nabla^\Gamma$ by

$$\Psi^\Gamma_\Lambda := \{\psi^\Gamma_\lambda : \lambda \in \Lambda\}, \quad \tilde{\Psi}^\Gamma_\Lambda := \{\tilde{\psi}^\Gamma_\lambda : \lambda \in \Lambda\}$$

finite subsets of the wavelet bases $\Psi, \tilde{\Psi}$ determined by the index set $\Lambda \subset \nabla$. Furthermore, for any (at most countable) collection $\Phi \subset L_2(\Gamma)$ let

$$S(\Phi) := \operatorname{clos}_{L_2}(\operatorname{span}(\Phi)).$$

Our objective is to find $u_\Lambda \in S(\Psi^\Gamma_\Lambda)$ such that

$$\langle \mathcal{L} u_\Lambda, v\rangle_\Gamma = \langle f, v\rangle_\Gamma, \quad v \in S(\Psi^\Gamma_\Lambda) \tag{5.1.3}$$

with the ansatz $u_\Lambda = \mathbf{d}^T_\Lambda \Psi^\Gamma_\Lambda$; this is equivalent to the discrete system

$$\langle \mathcal{L}\Psi^\Gamma_\Lambda, \Psi^\Gamma_\Lambda\rangle^T_\Gamma \mathbf{d}_\Lambda = \langle f, \Psi^\Gamma_\Lambda\rangle^T_\Gamma. \tag{5.1.4}$$

We will always assume that the Galerkin scheme is *stable* and explain next what this means in the present context. Associating with the pair $\Psi^\Gamma, \tilde{\Psi}^\Gamma$ the projectors

$$Q_\Lambda v := \langle v, \tilde{\Psi}^\Gamma_\Lambda\rangle_\Gamma \Psi^\Gamma_\Lambda, \quad Q^*_\Lambda v := \langle v, \Psi^\Gamma_\Lambda\rangle_\Gamma \tilde{\Psi}^\Gamma_\Lambda,$$

(which are adjoints of each other) stability means that

$$\|Q^*_\Lambda \mathcal{L} v\|_{H^{-t}(\Gamma)} \sim \|v\|_{H^t(\Gamma)}, \quad v \in S(\Psi^\Gamma_\Lambda),\ \#\Lambda \to \infty. \tag{5.1.5}$$

Of course, when $\mathcal{L}$ satisfies (5.1.1) and is self-adjoint, respectively, *strongly elliptic*, (5.1.5) is trivially satisfied. For sufficient conditions when $\mathcal{L}$ is a pseudodifferential operator, see, e.g., [19] and the literature cited there.

5.2. Preconditioning. As before we will assume in the following that Assumption A holds for $\mathcal{F} = H^s$, $|t| \le s < s_\Gamma$. The relations in Theorem 3.3 then read

$$\|v\|_{H^t(\Gamma)} \sim \|\langle v, \tilde{\Psi}^\Gamma\rangle_\Gamma \mathbf{D}^t\|_{\ell_2(\nabla^\Gamma)}, \quad \|v\|_{H^{-t}(\Gamma)} \sim \|\langle v, \Psi^\Gamma\rangle_\Gamma \mathbf{D}^{-t}\|_{\ell_2(\nabla^\Gamma)}, \tag{5.2.1}$$

where

$$\mathbf{D}^t := (2^{t|\lambda|}\delta_{\lambda,\lambda'})_{\lambda,\lambda' \in \nabla^\Gamma}.$$

Recall that analogous relations hold for the roles of Ψ and $\tilde{\Psi}$ interchanged. By $\mathbf{D}^s_\Lambda$ we denote the principal sections of $\mathbf{D}^s$ determined by Λ. It is well known that the norm equivalences (5.2.1) together with the stability of the Galerkin scheme implies the following fact [19, 16].

THEOREM 5.1. *Under the above assumptions one has*

$$\operatorname{cond}_2\left(\mathbf{D}^{-t}_\Lambda \langle \mathcal{L}\Psi^\Gamma_\Lambda, \Psi^\Gamma_\Lambda\rangle_\Gamma \mathbf{D}^{-t}_\Lambda\right) = \mathcal{O}(1), \quad \#\Lambda \to \infty, \tag{5.2.2}$$

where $\operatorname{cond}_2(\mathbf{B}) := \|\mathbf{B}\|\|\mathbf{B}^{-1}\|$ *denotes the spectral condition number of the matrix* $\mathbf{B}$.

Thus diagonal scalings of stiffness matrices relative to the above wavelet bases produce well-conditioned system matrices so that *iterative solvers* have a chance to work with asymptotically optimal efficiency; see [16].

5.3. Computation of stiffness matrices. Let us briefly indicate next what it means concretely to compute the stiffness matrices

$$\mathbf{A}_\Lambda := \langle \mathcal{L}\Psi^\Gamma_\Lambda, \Psi^\Gamma_\Lambda \rangle^T_\Gamma.$$

Throughout this section we will assume that the projections P_i are defined with respect to the localized extensions (4.3.4) constructed in section 4.3. Let $\lambda = (i, \nu), \lambda' = (l, \mu) \in \nabla^\Gamma$. It is instructive to see first how an inner product with a wavelet reduces to an inner product over $\Box$. Of course, when $\nu \in \Lambda^\uparrow_i$ (see (4.3.3)) one simply obtains, in view of (3.3.3) and the fact that E_i reduces in this case to multiplication with χ_{Γ_i},

$$\langle v, \psi^\Gamma_\lambda \rangle_\Gamma = \langle v, \psi^{\Gamma_i,\downarrow}_\nu \rangle_{\Gamma_i} = \langle |\partial\kappa_i| v \circ \kappa_i, \psi^{\mathbf{Z}^{(i)}}_\nu \rangle_\Box, \tag{5.3.1}$$

where we have used (2.2.4) in the last step. When $\nu \notin \Lambda^\uparrow_i$ so that $E_i \psi^{\Gamma_i,\downarrow}_\nu$ reaches into $\Gamma^\uparrow_i \setminus \Gamma_i$ a natural option is to compute each contribution of the involved patches according to

$$\begin{aligned}\langle v, \psi^\Gamma_\lambda \rangle_\Gamma &= \langle v, \psi^{\Gamma_i,\downarrow}_\nu \rangle_{\Gamma_i} + \sum_{\Gamma_l \subset \Gamma^\uparrow_i \setminus \Gamma_i} \langle v, E_i \psi^{\Gamma_i,\downarrow}_\nu \rangle_{\Gamma_l} \\ &= \langle |\partial\kappa_i| v \circ \kappa_i, \psi^{\mathbf{Z}^{(i)}}_\nu \rangle_\Box + \sum_{\Gamma_l \subset \Gamma^\uparrow_i \setminus \Gamma_i} \langle |\partial\kappa_l| v \circ \kappa_l, (E_i \psi^{\Gamma_i,\downarrow}_\nu) \circ \kappa_l \rangle_\Box. \end{aligned} \tag{5.3.2}$$

Suppose now that the extensions are defined by lifting (4.1.6). Note that the patches $\Gamma_l \subset \Gamma^\uparrow_i$ induce a partition of $\Box_{\tilde{\mathbf{Z}}^{(i)}}$ consisting of the patches $\rho_{i,l}(\Box)$, $\Gamma_l \subset \Gamma^\uparrow_i \setminus \Gamma_i$, so that $\Gamma_l = \kappa^\uparrow_i \circ \rho_{i,l}(\Box)$, i.e., $\kappa_l = \kappa^\uparrow_i \circ \rho_{i,l}$. Using these relations and substituting (4.1.6), (5.3.2) becomes

$$\begin{aligned}\langle v, \psi^\Gamma_\lambda \rangle_\Gamma = {} & \langle |\partial\kappa_i| v \circ \kappa_i, \psi^{\mathbf{Z}^{(i)}}_\nu \rangle_\Box \\ & + \sum_{\Gamma_l \subset \Gamma^\uparrow_i \setminus \Gamma_i} \langle |\partial\kappa_l| v \circ \kappa_l, (A_{\tilde{\mathbf{Z}}^{(i)}} \psi^{\mathbf{Z}^{(i)}}_\nu) \circ \rho_{i,l} \rangle_\Box. \end{aligned} \tag{5.3.3}$$

An alternative is to use the adjoints to obtain via (2.2.4) and (3.3.3)

$$\begin{aligned}\langle v, \psi^\Gamma_\lambda \rangle_\Gamma &= \langle v, \chi_{\Gamma^\uparrow_i} E_i \psi^{\Gamma_i,\downarrow}_\lambda \rangle_\Gamma = \langle E^*_i(v\,|_{\Gamma^\uparrow_i}), \psi^{\Gamma_i,\downarrow}_\nu \rangle_{\Gamma_i} \\ &= \langle |\partial\kappa_i| \kappa^*_i \left(E^*_i (v\,|_{\Gamma^\uparrow_i}) \right), \psi^{\mathbf{Z}^{(i)}}_\nu \rangle_\Box = \langle A^*_{\tilde{\mathbf{Z}}^{(i)}} \left(|\partial\kappa^\uparrow_i| v \circ \kappa^\uparrow_i \right), \psi^{\mathbf{Z}^{(i)}}_\nu \rangle_\Box, \end{aligned} \tag{5.3.4}$$

where we have used (4.1.7).

Which option is preferable depends on the particular nature of the extensions and, since we are particularly interested in the entries of stiffness matrices which means that the above expressions have to be applied to $v = \mathcal{L}\psi^\Gamma_{\lambda'}$, also on the nature of $\mathcal{L}$.

Let us briefly discuss here the case

$$\mathcal{L}v = \int_\Gamma K(\cdot, y) v(y) ds_y. \tag{5.3.5}$$

Substituting $v = \mathcal{L}\psi^\Gamma_\lambda$ in (5.3.4) yields the following relations.

REMARK 5.3.1. *Under the above hypotheses one has for $\lambda = (i, \nu)$, $\lambda' = (l, \mu)$*

$$\langle \mathcal{L}\psi^\Gamma_{\lambda'}, \psi^\Gamma_\lambda \rangle_\Gamma = \int_\Box \int_\Box L_{l,i}(x, y) \psi^{\mathbf{Z}^{(l)}}_\mu (y) \psi^{\mathbf{Z}^{(i)}}_\nu (x) dx dy, \tag{5.3.6}$$

where the kernel $L_{l,i}$ *has the following form:*

$$L_{l,i}(x,y) = \begin{cases} |\partial\kappa_i(x)||\partial\kappa_l(y)|K(\kappa_i(x),\kappa_l(y)), \\ \qquad \nu\in\Lambda_i^{\uparrow}, \mu\in\Lambda_l^{\uparrow}; \\ (A^*_{\tilde{\mathbf{Z}}^{(i)}}\otimes I)\left(|\partial\kappa_i^{\uparrow}(x)||\partial\kappa_l(y)|\,K(\kappa_i^{\uparrow}(x),\kappa_l(y))\right), \\ \qquad \nu\in\nabla^i\setminus\Lambda_i^{\uparrow}, \mu\in\Lambda_l^{\uparrow}; \\ (I\otimes A^*_{\tilde{\mathbf{Z}}^{(l)}})\left(|\partial\kappa_i(x)||\partial\kappa_l^{\uparrow}(y)|\,K(\kappa_i(x),\kappa_l^{\uparrow}(y))\right), \\ \qquad \nu\in\Lambda_i^{\uparrow}, \mu\in\nabla^l\setminus\Lambda_l^{\uparrow}; \\ (A^*_{\tilde{\mathbf{Z}}^{(i)}}\otimes A^*_{\tilde{\mathbf{Z}}^{(l)}})\left(|\partial\kappa_i^{\uparrow}(x)||\partial\kappa_l^{\uparrow}(y)|\,K(\kappa_i^{\uparrow}(x),\kappa_l^{\uparrow}(y))\right), \\ \qquad \mu\in\nabla^l\setminus\Lambda_l^{\uparrow}, \nu\in\nabla^i\setminus\Lambda_i^{\uparrow}. \end{cases} \tag{5.3.7}$$

Clearly the smoothness of the modified kernel and consequently the compressibility of corresponding stiffness matrices depend on the regularity of the extended parametrization $\kappa_i^{\uparrow}$, recall Property G.

5.4. Matrix compression. Suppose now that the kernel K in (5.3.5) is smooth everywhere except for $x=y$ and satisfies the estimates

$$|\partial_x^{\alpha}\partial_y^{\beta}K(x,y)| \lesssim \operatorname{dist}(x,y)^{-(n+2t+|\alpha|+|\beta|)} \tag{5.4.1}$$

for any $\alpha,\beta\in\mathbb{N}_0^d$ such that $n+2t+|\alpha|+|\beta|>0$ with constants depending on α,β; see, e.g., [16, 19, 45] for the background of such estimates. The representation (5.3.7) has the following important consequence with regard to *matrix compression.*

THEOREM 5.2. *Let* Ω_λ *denote the support of* ψ_λ^{Γ}. *Under the above hypotheses one has the estimates*

$$2^{-(|\lambda'|+|\lambda|)t}|\langle\mathcal{L}\psi_{\lambda'}^{\Gamma},\psi_\lambda^{\Gamma}\rangle_\Gamma| \lesssim \frac{2^{-||\lambda|-|\lambda'||\sigma}}{(1+2^{\min(|\lambda|,|\lambda'|)}\operatorname{dist}(\Omega_\lambda,\Omega_{\lambda'}))^{n+2\tilde{d}+2t}}, \tag{5.4.2}$$

provided that when $2^{\min(|\lambda|,|\lambda'|)}\operatorname{dist}(\Omega_\lambda,\Omega_{\lambda'})\gtrsim 1$ *the kernel* $L_{l,i}(x,y)$ *has bounded* $\tilde{d}$*th order derivatives in* x *and* y *on* $\operatorname{supp}(\psi_\nu^{\mathbf{Z}^{(i)}})\times\operatorname{supp}(\psi_\mu^{\mathbf{Z}^{(l)}})$, $\lambda=(i,\nu)$, $\lambda'=(l,\mu)$. *Here* $\tilde{d}$ *is the integer from Properties B* (iii) *and* $\sigma>\frac{d}{2}$ *depends on the regularity of the wavelets.*

Proof. We sketch only the argument and refer to [23] for a detailed discussion of inequalities of the above type. Recall from (3.1.11) that

$$\langle P,\psi_\lambda^{\mathbf{Z}^{(i)}}\rangle_\square = 0 \quad \forall\ P\in\Pi_{\tilde{d}}(\square)\cap\mathcal{F}(\square)_{\tilde{\mathbf{Z}}^{(i)}},\ \lambda\in\nabla_-^{\mathbf{Z}^{(i)}}. \tag{5.4.3}$$

This, in turn, is easily seen to imply that

$$|\langle f,\psi_\lambda^{\mathbf{Z}^{(i)}}\rangle_\square| \lesssim 2^{-|\lambda|(\tilde{d}+\frac{n}{2})}\|f\|_{H_\infty^{\tilde{d}}(\Omega_\lambda^\square)}, \quad \lambda\in\nabla_-^{\mathbf{Z}^{(i)}}, \tag{5.4.4}$$

whenever

$$f\in\tilde{\mathcal{F}}(\square)_{\tilde{\mathbf{Z}}^{(i)}}\cap C^{\tilde{d}}(\overline{\Omega}_\lambda^\square). \tag{5.4.5}$$

Whenever the extensions are sufficiently smooth the kernel $L_{l,i}(\cdot,y)$ belongs for every y as a function of x, due to the action of the restriction $A^*_{\tilde{\mathbf{Z}}^{(i)}}$, to the space $\tilde{\mathcal{F}}(\square)_{\tilde{\mathbf{Z}}^{(i)}}$, which by the above remarks provides an estimate of type (5.4.4). Applying (5.4.4) successively with respect to each variable x and y one deduces for $\operatorname{dist}(\Omega_\lambda, \Omega_{\lambda'}) \gtrsim 2^{-|\lambda'|}$ where $|\lambda'| \leq |\lambda|$ the estimate

$$|\langle \mathcal{L}\psi^{\Gamma}_{\lambda'}, \psi^{\Gamma}_{\lambda}\rangle_{\Gamma}| \lesssim \frac{2^{-(|\lambda|+|\lambda'|)(n/2+\tilde{d})}}{(\operatorname{dist}(\Omega_\lambda, \Omega_{\lambda'}))^{n+2\tilde{d}+2t}};$$

see [21, 24] for details on the last step. Using a Cauchy–Schwarz argument for the remaining cases (see [13]) one finally obtains (5.4.2). □

Recall that, due to the boundary conditions of the wavelet bases on □, the moment conditions in (5.4.3) are *constrained.* However, we stress that, on account of (5.3.7) and the properties of the E_i, E_i^*, the modified kernels satisfy exactly the right constraints so that in smooth regions the cancellation properties hold *patchwise* with maximum order which distinguishes the present concept from those based on gluing neighboring bases directly [7, 22].

5.5. Domain decomposition. The above discussion already indicates that actual computations are essentially reduced to the unit cube □. It is then natural to extend this conceptually to the solution process which is the somewhat different point of view to be briefly discussed next. Obviously, (5.1.2) is equivalent to

$$V\mathcal{L}S(Tu) = Vf,$$

where T, S, V are defined by (2.4.6) and (2.4.7). Thus setting

$$\Pi_\downarrow := \prod_{i=1}^{M} H^s(\Gamma_i)^\downarrow, \quad \Pi_\uparrow^* := \prod_{i=1}^{M} H^{-s}(\Gamma_i)^\uparrow,$$

one has

$$\mathcal{L}_\Pi := V\mathcal{L}S : \Pi_\downarrow \to \Pi_\uparrow^*, \quad \mathbf{f} := Vf \in \Pi_\uparrow^*, \tag{5.5.1}$$

and we wish to find for every $\mathbf{f} \in \Pi_\uparrow^*$ the unique $\mathbf{u} \in \Pi_\downarrow$ such that

$$\mathcal{L}_\Pi \mathbf{u} = \mathbf{f}. \tag{5.5.2}$$

Of course, given $\mathbf{u}$ satisfying (5.5.2), $u = S\mathbf{u}$ solves (5.1.2). One easily confirms that

$$\mathcal{L}_\Pi = (\mathcal{L}_{i,l})_{i,l=1}^{M}, \quad \mathcal{L}_{i,l}w := (P_i^* \mathcal{L} P_l \chi_{\Gamma_l} w)\,|_{\Gamma_i}\,.$$

Moreover, recalling from (3.3.1) the inner product

$$\langle \{v_i\}, \{u_i\}\rangle_\Pi := \sum_{i=1}^{M} \langle v_i, u_i\rangle_{\Gamma_i}$$

on the product space, one has by definition

$$\langle \mathcal{L}_\Pi\{u_i\}, \{v_i\}\rangle_\Pi = \langle \mathcal{L}u, v\rangle_\Gamma, \quad \{u_i\} = Tu.$$

The weak formulation $\langle \mathcal{L}_\Pi \mathbf{u}, \mathbf{v}\rangle_\Pi = \langle \mathbf{f}, \mathbf{v}\rangle_\Pi$, $\mathbf{v} \in \Pi_\downarrow$, of (5.5.2) takes the form

$$\sum_{l=1}^{N} \left\langle \sum_{i=1}^{N} \mathcal{L}_{l,i} u_i - f_l, v_l \right\rangle_{\Gamma_l} = 0, \quad \mathbf{v} = \{v_i\}_{i=1}^{N} \in \Pi_\downarrow. \tag{5.5.3}$$

A natural idea is to solve (5.5.3) iteratively. The simplest version is a Jacobi-type scheme yielding $\mathbf{u}^n$ by

$$\mathcal{L}_{i,i}u_i^{n+1} = f_i - \sum_{l\neq i}\mathcal{L}_{i,l}u_l^n, \quad i=1,\dots,N. \tag{5.5.4}$$

Let us briefly discuss now the convergence properties of the iteration (5.5.4) (or a corresponding relaxation version). First note that the ellipticity (5.1.1) combined with Theorem 2.3 yield

$$\|\mathcal{L}_\Pi \mathbf{v}\|_{\Pi_\uparrow^*} \sim \|\mathbf{v}\|_{\Pi_\downarrow}, \quad \mathbf{v}\in\Pi_\downarrow.$$

Specifically, when $\mathcal{L}$ is self-adjoint and positive definite this means

$$\|\mathbf{v}\|_{\Pi_\downarrow}^2 \sim \langle\mathcal{L}_\Pi\mathbf{v},\mathbf{v}\rangle_\Pi, \quad \mathbf{v}\in\Pi_\downarrow.$$

which, in particular, implies that

$$\|\mathcal{L}_{i,i}v\|_{H^{-t}(\Gamma_i)^\uparrow} \sim \|v\|_{H^t(\Gamma_i)^\downarrow}, \quad i=1,\dots,N. \tag{5.5.5}$$

Hence, the *local problems*

$$\mathcal{L}_{i,i}u_i = g_i, \quad i=1,\dots,N, \tag{5.5.6}$$

on □ are elliptic. Of course, the obvious analog to Theorem 5.1, which holds on account of norm equivalences of the type (3.1.17) on □ (for $\mathcal{F}=H^t$), ensures that these local problems (5.5.6) can, in turn, be solved iteratively with asymptotically optimal complexity.

Remark 5.5.1. *Thus we see that the stability properties of the local wavelet bases combined with Theorem* 2.3 *lead to* stable splittings *in the sense of* [41] *which means that convergence is guaranteed for a controlled number of outer and inner iterations. We emphasize that this is true* regardless *of whether $\mathcal{L}$ is a (local) differential operator or a (global) singular integral operator as long as* (5.1.1) *holds. Roughly speaking, the localization properties of wavelet bases makes differential and integral operators equally tractable by domain decomposition schemes. In combination with the compression estimates in section* 5.4 *this is expected to give rise to numerical schemes by which the problem can be solved within the accuracy admitted by the discretization error at an expense of CPU and storage that remains proportional to the problem size while in addition parallel techniques are naturally incorporated. Details will be given in* [23].

We also remark that the global discretization discussed in section 5.1 as well as the local problems (5.5.6) satisfy all assumptions required in [13] for the analysis of *adaptive* strategies. In particular, the above domain decomposition concept offers a natural marriage between parallel techniques and adaptive strategies based on the wavelet bases $\Psi^{\Gamma_i,\downarrow}\subset H^t(\Gamma_i)^\downarrow$ applied independently to each local problem.

Appendix A: Proof of Proposition 2.4.1.

ad (i):. First consider the case $i<j$. Since by (2.3.15) $\Gamma_i\cap\Gamma_j^\uparrow=\emptyset$, we infer from (2.4.2) for $i=1$ and (2.4.5) that

$$P_1P_jv = \chi_{\Gamma_1^\uparrow}E_1\left((P_jv)\,|_{\Gamma_1}\right) = \chi_{\Gamma_1^\uparrow}E_1(0)=0.$$

Suppose that (2.4.8) holds for $l<i<j$. Again by (2.3.15), the definition (2.4.3) and (2.4.5) the induction assumption gives

$$P_iP_jv = \chi_{\Gamma_i^\uparrow}E_i\left(\left(P_jv-\sum_{l<i}P_lP_jv\right)|_{\Gamma_i}\right) = \chi_{\Gamma_i^\uparrow}E_i\left((P_jv)\,|_{\Gamma_i}\right)=0,$$

which confirms that

$$P_iP_j = 0, \quad i < j. \tag{5.5.7}$$

To prove the rest of (2.4.8) we use (2.4.2) to conclude first that

$$P_1^2 v = \chi_{\Gamma_1^\uparrow} E_1 \left((\chi_{\Gamma_1^\uparrow} E_1(v\,|_{\Gamma_1}))\,|_{\Gamma_1} \right) = \chi_{\Gamma_1^\uparrow} E_1\,(v\,|_{\Gamma_1}) = P_1 v, \tag{5.5.8}$$

while for $i > 1$, in view of (5.5.8),

$$P_iP_1v = \chi_{\Gamma_i^\uparrow} E_i \left(\left(P_1v - \sum_{j<i} P_jP_1v \right) |_{\Gamma_i} \right) = \chi_{\Gamma_i^\uparrow} E_i \left(\left(- \sum_{1<j<i} P_jP_1v \right) |_{\Gamma_i} \right). \tag{5.5.9}$$

Thus when $i = 2$ the right-most term in (5.5.9) is vacuous so that $P_2P_1v = 0$ while induction based on (5.5.9) easily yields

$$P_iP_1 = 0, \quad i > 1. \tag{5.5.10}$$

In view of (5.5.8) and (5.5.10), we may assume now that

$$P_jP_\ell = \delta_{j,\ell}P_j, \quad j < i,\ \ell \le j \tag{5.5.11}$$

and

$$P_iP_r = 0, \quad r < \ell. \tag{5.5.12}$$

To advance the induction assumptions let us verify first that $P_i^2 = P_i$. To this end, note first that in analogy to (5.5.8)

$$\chi_{\Gamma_i^\uparrow} E_i \left((\chi_{\Gamma_i^\uparrow} E_i(w\,|_{\Gamma_i}))\,|_{\Gamma_i} \right) = \chi_{\Gamma_i^\uparrow} E_i(w\,|_{\Gamma_i}). \tag{5.5.13}$$

Now (5.5.7) yields

$$\begin{aligned} P_i^2 v &= \chi_{\Gamma_i^\uparrow} E_i \left(\left(P_iv - \sum_{j<i} P_jP_iv \right) |_{\Gamma_i} \right) = \chi_{\Gamma_i^\uparrow} E_i((P_iv)\,|_{\Gamma_i}) \\ &= \chi_{\Gamma_i^\uparrow} E_i \left(\chi_{\Gamma_i^\uparrow} E_i \left(\left(v - \sum_{j<i} P_jv \right) |_{\Gamma_i} \right) |_{\Gamma_i} \right) = \chi_{\Gamma_i^\uparrow} E_i \left(\left(v - \sum_{j<i} P_jv \right) |_{\Gamma_i} \right) = P_iv, \end{aligned}$$

where we have used (5.5.13) in the second but last step. Furthermore, by (5.5.7) and (5.5.11),

$$\begin{aligned} P_iP_\ell v &= \chi_{\Gamma_i^\uparrow} E_i \left(\left(P_\ell v - \sum_{j<i} P_jP_\ell v \right) |_{\Gamma_i} \right) = \chi_{\Gamma_i^\uparrow} E_i \left(\left(P_\ell v - P_\ell v - \sum_{j=\ell+1}^{i-1} P_jP_\ell v \right) |_{\Gamma_i} \right) \\ &= \chi_{\Gamma_i^\uparrow} E_i(0) = 0, \quad \text{for} \quad \ell < i, \end{aligned}$$

which advances the induction and proves (i).

ad (ii):. Let $h := v - \sum_{j=1}^{N} P_j v$. By (2.4.8) one has

$$P_i h = P_i v - \sum_{j=1}^{N} P_i P_j v = P_i v - P_i v = 0,$$

$i = 1, \ldots, N$. Thus

$$0 = P_i h = \chi_{\Gamma_i^{\uparrow}} E_i \left(\left(h - \sum_{j<i} P_j h \right) |_{\Gamma_i} \right) = \chi_{\Gamma_i^{\uparrow}} E_i \left(h \,|_{\Gamma_i} \right)$$

implies $h = 0$ which confirms (ii).

ad (iii):. The first relation in (2.4.10) is an immediate consequence of (2.3.15) and the definition of P_i. Likewise for $j > i$, it is clear that

$$P_1 \chi_{\Gamma_j} v = \chi_{\Gamma_1^{\uparrow}} E_1 \left((\chi_{\Gamma_j} v) \,|_{\Gamma_1} \right) = \chi_{\Gamma_1^{\uparrow}} E_1(0) = 0,$$

and assuming that $P_\ell \chi_{\Gamma_j} v = 0$, $\ell < i < j$, one has

$$P_i(\chi_{\Gamma_j} v) = \chi_{\Gamma_i^{\uparrow}} E_i \left(\left(\chi_{\Gamma_j} v - \sum_{\ell<i} P_\ell \chi_{\Gamma_j} v \right) |_{\Gamma_i} \right) = \chi_{\Gamma_i^{\uparrow}} E_i \left((\chi_{\Gamma_j} v) \,|_{\Gamma_i} \right) = \chi_{\Gamma_i^{\uparrow}} E_i(0) = 0,$$

which proves (iii) and finishes the proof of the proposition. □

Appendix B: Proof of Theorem 2.2. If $v \,|_{\Gamma_i} \in \mathcal{F}(\Gamma_i)^{\downarrow}$, then $\hat{v} := \chi_{\Gamma_i^{\uparrow\uparrow}} v = \sum_{l=i}^{N} \chi_{\Gamma_l} v \in \mathcal{F}(\Gamma)$. By (2.4.9), one has

$$\begin{aligned} \hat{v} \,|_{\Gamma_i} &= \left(\sum_{j=1}^{N} P_j \hat{v} \right) |_{\Gamma_i} = \left(\sum_{j=1}^{N} \chi_{\Gamma_i} P_j \hat{v} \right) |_{\Gamma_i} = \left(\sum_{j=1}^{i} P_j \hat{v} \right) |_{\Gamma_i} \\ &= \left(\sum_{j=1}^{i} \sum_{l=i}^{N} P_j \chi_{\Gamma_l} v \right) |_{\Gamma_i} = \left(P_i \sum_{l=i}^{N} \chi_{\Gamma_l} v \right) |_{\Gamma_i} = (P_i \hat{v}) \,|_{\Gamma_i}, \end{aligned}$$

which proves the first relation in (2.5.13). Employing (2.4.13) and (2.4.14), the argument for the second relation in (2.5.13) is completely analogous.

By definition (2.4.2) and Assumption A (2.5.7), we know that

$$\begin{aligned} \|P_1 v\|_{\mathcal{F}(\Gamma_1)^{\uparrow}} &= \|E_1(v \,|_{\Gamma_1})\|_{\mathcal{F}(\Gamma_1^{\uparrow})} = \|E_1(v \,|_{\Gamma_1})\|_{\mathcal{F}(\Gamma_1^{\uparrow})^{\downarrow}} \\ &\lesssim \|v\|_{\mathcal{F}(\Gamma_1)^{\downarrow}} = \|v\|_{\mathcal{F}(\Gamma_1)} \leq \|v\|_{\mathcal{F}(\Gamma)}, \end{aligned}$$

where we have used several times that Γ_1 has no inflow boundary. This confirms the first relations in (2.5.16) and (2.5.15) for $i = 1$. Now suppose that the first relations in (2.5.15) and (2.5.16) have been verified for $1 \leq \ell < i$. Thus

$$h_i := \left(v - \sum_{j<i} P_j v \right) \in \mathcal{F}(\Gamma)$$

and for any $\ell < i$ one has by (2.4.9) and (2.4.10),

$$\begin{aligned} \chi_{\Gamma_\ell} h_i &= \sum_{r=1}^{N} \chi_{\Gamma_\ell} P_r h_i = \sum_{r=1}^{\ell} \chi_{\Gamma_\ell} P_r h_i \\ &= \sum_{r=1}^{\ell} \chi_{\Gamma_\ell} \left(P_r v - \sum_{j<i} P_r P_j v \right) = \sum_{r=1}^{\ell} \chi_{\Gamma_\ell} (P_r v - P_r v) = 0, \end{aligned} \tag{5.5.14}$$

where we have used (2.4.8) of Proposition 2.4.1 in the last step. Moreover, since $h_i = \chi_{\Gamma_i^{\uparrow\uparrow}} h_i$, (2.5.13) and (2.4.8) yield

$$h_i = P_i h_i = P_i v \in \mathcal{F}(\Gamma). \tag{5.5.15}$$

Therefore (5.5.14) also says that $(P_i v) \mid_{\Gamma_i} \in \mathcal{F}(\Gamma_i)^{\downarrow}$ which confirms together with (2.5.13) the first relation in (2.5.14).

Thus, by (2.5.7),

$$\begin{aligned} \|P_i v\|_{\mathcal{F}(\Gamma)} &= \|\chi_{\Gamma_i^{\uparrow}} E_i(h_i \mid_{\Gamma_i})\|_{\mathcal{F}(\Gamma)} = \|E_i(h_i \mid_{\Gamma_i})\|_{\mathcal{F}(\Gamma_i^{\uparrow})^{\downarrow}} \\ &\lesssim \|h_i \mid_{\Gamma_i} \|_{\mathcal{F}(\Gamma_i)^{\downarrow}} \leq \|h_i\|_{\mathcal{F}(\Gamma)} \leq \|v - \sum_{j<i} P_j v\|_{\mathcal{F}(\Gamma)} \\ &\leq \|v\|_{\mathcal{F}(\Gamma)} + \sum_{j<i} \|P_j v\|_{\mathcal{F}(\Gamma)} \lesssim \|v\|_{\mathcal{F}(\Gamma)}, \end{aligned}$$

which, by induction, confirms the first relation in (2.5.16) $\forall i = 1, \dots N$. Since by definition,

$$\|(P_i v) \mid_{\Gamma_i} \|_{\mathcal{F}(\Gamma_i)^{\downarrow}} = \|P_i v\|_{\mathcal{F}(\Gamma_i^{\downarrow})} \leq \|P_i v\|_{\mathcal{F}(\Gamma)};$$

also the first relation in (2.5.15) and hence all claims concerning the projectors P_i have been verified.

As for the adjoints P_i^*, we have by (2.4.11), (2.4.13), and (2.4.14),

$$P_i^* w = \left(I - \sum_{j<i} P_j^* \right) \chi_{\Gamma_i} E_i^*(w \mid_{\Gamma_i^{\uparrow}}) = \sum_{j=i}^{N} P_j^* \chi_{\Gamma_i} E_i^*(w \mid_{\Gamma_i^{\uparrow}}) = P_i^* \chi_{\Gamma_i} E_i^*(w \mid_{\Gamma_i^{\uparrow}}). \tag{5.5.16}$$

For $r \geq i$ one obtains by (2.4.14), that $\chi_{\Gamma_r} P_j^* = 0$ whenever $j < i$, so that for w replaced for simplicity by $w \mid_{\Gamma_i^{\uparrow}}$

$$\chi_{\Gamma_r} \sum_{j<i} P_j^* \chi_{\Gamma_i} E_i^* w = 0, \quad r \geq i. \tag{5.5.17}$$

Hence we infer from (2.4.14) and (5.5.16) that

$$\chi_{\Gamma_r} P_i^* w = 0, \quad r > i, \quad \chi_{\Gamma_i} P_i^* w = \chi_{\Gamma_i} E_i^* w, \tag{5.5.18}$$

which, by (2.5.8), means that

$$P_i^* w \in \tilde{\mathcal{F}}(\Gamma_i)^{\uparrow} \tag{5.5.19}$$

and

$$\|(P_i^* w)\,|_{\Gamma_i^{\uparrow\uparrow}}\,\|_{\tilde{\mathcal{F}}(\Gamma_i^{\uparrow\uparrow})} \;\lesssim\; \|w\,|_{\Gamma_i^{\uparrow\uparrow}}\,\|_{\tilde{\mathcal{F}}(\Gamma_i^{\uparrow\uparrow})}. \tag{5.5.20}$$

Therefore

$$\|(P_i^* v)\,|_{\Gamma_i}\|_{\tilde{\mathcal{F}}(\Gamma_i)^\uparrow} \;\lesssim\; \|v\|_{\tilde{\mathcal{F}}(\Gamma_i^\uparrow)}, \quad v \in \tilde{\mathcal{F}}(\Gamma_i^\uparrow),\; i = 1, \ldots, N, \tag{5.5.21}$$

whence the second parts of (2.5.14) and (2.5.15) follow.

It remains to verify the second part of (2.5.16). It suffices to show that the P_i are bounded in $\tilde{\mathcal{F}}^*(\Gamma)$. To this end note first that it follows from (2.5.8) that

$$\|E_i w\|_{\tilde{\mathcal{F}}^*(\Gamma_i^\uparrow)} \;\lesssim\; \|w\|_{\tilde{\mathcal{F}}^*(\Gamma_i)^\downarrow}, \quad w \in \tilde{\mathcal{F}}^*(\Gamma_i)^\downarrow. \tag{5.5.22}$$

In fact,

$$\begin{aligned}
\|E_i w\|_{\tilde{\mathcal{F}}^*(\Gamma_i^\uparrow)} &= \sup_{\|v\|_{\tilde{\mathcal{F}}(\Gamma_i^\uparrow)}=1} |\langle E_i w, v\rangle_{\Gamma_i^\uparrow}| = \sup_{\|v\|_{\tilde{\mathcal{F}}(\Gamma_i^\uparrow)}=1} |\langle w, E_i^* v\rangle_{\Gamma_i}| \\
&\le \sup_{\|v\|_{\tilde{\mathcal{F}}(\Gamma_i^\uparrow)}=1} \|w\|_{\tilde{\mathcal{F}}^*(\Gamma_i)^\downarrow} \|E_i^* v\|_{\tilde{\mathcal{F}}(\Gamma_i)^\uparrow} \\
&\lesssim \sup_{\|v\|_{\tilde{\mathcal{F}}(\Gamma_i^\uparrow)}=1} \|w\|_{\tilde{\mathcal{F}}^*(\Gamma_i)^\downarrow} \|v\|_{\mathcal{F}(\tilde{\Gamma}_i^\uparrow)},
\end{aligned}$$

where we have used (2.5.8) in the last step. Now for $i = 1$ the boundedness of P_i in $\tilde{\mathcal{F}}^*(\Gamma)$ follows immediately from the definition and (5.5.22). Suppose that the P_j are bounded in $\tilde{\mathcal{F}}^*(\Gamma)$ for $1 \le j < i$. Thus for $v \in \tilde{\mathcal{F}}^*(\Gamma)$ the distribution $h_i = (v - \sum_{j<i} P_j v)$ belongs to $\tilde{\mathcal{F}}^*(\Gamma)$ and $\|h_i\|_{\tilde{\mathcal{F}}^*(\Gamma)} \lesssim \|v\|_{\tilde{\mathcal{F}}^*(\Gamma)}$. From the beginning of the proof we know that $\chi_{\Gamma_l} h_i = 0$ for $l < i$ so that $\chi_{\Gamma_i^{\uparrow\uparrow}} h_i \in \tilde{\mathcal{F}}^*(\Gamma)$. The assertion follows then by (2.5.7) as before. □

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*. Pure and Applied Mathematics, Vol. 65, Academic Press, London, New York, 1978.

[2] B. Alpert, G. Beylkin, R. Coifman, and V. Rokhlin, *Wavelet-like bases for the fast solution of second-kind integral equations*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 159–184.

[3] A. Averbuch, G. Beylkin, R. Coifman, and M. Israeli, *Multiscale inversion of elliptic operators*, in Signal and Image Representation in Combined Spaces, J. Zeevi and R. Coifman, eds., Academic Press, San Diego, CA, 1995, pp. 1–16.

[4] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Springer-Verlag, Berlin, New York, 1976.

[5] G. Beylkin, R. R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms* I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[6] F. Bornemann, B. Erdmann, and R. Kornhuber, *A posteriori error estimates for elliptic problems in two and three space dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 1188–1204.

[7] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[8] J. M. Carnicer, W. Dahmen, and J. M. Peña, *Local decomposition of refinable spaces*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 127–153.

[9] Z. Ciesielski and T. Figiel, *Spline bases in classical function spaces on compact C^∞ manifolds, Parts* I *and* II, Studia Math., 76 (1983), pp. 1–58; 95–136.

[10] A. Cohen, I. Daubechies, and P. Vial, *Wavelets on the interval and fast wavelet transforms*, Appl. Comput. Harmon. Anal., 1 (1993), pp. 54–81.

[11] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive Wavelet Methods for Elliptic Operator Equations—Convergence rates*, IGPM Report # 165, RWTH Aachen, 1998.

[12] A. Cohen and R. Masson, *Wavelet Adaptive Methods for Second Order Elliptic Problems, Boundary Conditions, and Domain Decomposition*, preprint, 1997.

[13] S. Dahlke, W. Dahmen, R. Hochmuth, and R. Schneider, *Stable multiscale bases and local error estimation for elliptic problems*, Appl. Numer. Math., 23 (1997), pp. 21–47.

[14] W. Dahmen, *Stability of multiscale transformations*, J. Fourier Anal. Appl., 2 (1996), pp. 341–361.

[15] W. Dahmen, *Multiscale analysis, approximation, and interpolation spaces*, in Approximation Theory VIII, Wavelets and Multilevel Approximation, C.K. Chui and L.L. Schumaker, eds., World Scientific, River Edge, NJ, 1995, pp. 47–88.

[16] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228, Cambridge University Press, Cambridge, UK, 1197.

[17] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.

[18] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal spline wavelets on the interval—stability and moment conditions*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 132–196.

[19] W. Dahmen, S. Prössdorf, and R. Schneider, *Multiscale methods for pseudo-differential equations on smooth manifolds*, in Proceedings of the International Conference on Wavelets: Theory, Algorithms, and Applications, C.K. Chui, L. Montefusco, and L. Puccio, eds., Academic Press, San Diego, 1994, pp. 385–424.

[20] W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential equations* II: *Matrix compression and fast solution*, Adv. Comput. Math., 1 (1993), pp. 259–335.

[21] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., in press.

[22] W. Dahmen and R. Schneider, *Wavelets with complementary boundary conditions—function spaces on the cube*, Results Math., 34 (1998), pp. 255–293.

[23] W. Dahmen and R. Schneider, *Wavelets on manifolds—Application to boundary integral equations*, in preparation.

[24] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[25] R. A. DeVore, B. Jawerth, and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math., 114 (1992), pp. 737–785.

[26] R. A. DeVore and V. A. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[27] J. A. Gregory, V. K. H. Lau, and J. Zhou, *Smooth parametric surfaces and n-sided patches*, in Computation of Curves and Surfaces, W. Dahmen, M. Gasca, and C.A. Micchelli, eds., NATO ASI Series, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1990.

[28] M. Griebel and P. Oswald, *Remarks on the abstract theory of additive and multiplicative Schwarz algorithms*, Numer. Math., 70 (1995), pp. 163–180.

[29] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[30] J. Hahn, *Geometric continuous patch complexes*, Comput. Aided Geom. Design, 6 (1989), pp. 55–67.

[31] K. Höllig and H. Mögerle, *G-splines*, Comput. Aided Geom. Design, 7 (1990), pp. 197–207.

[32] S. Jaffard, *Wavelet methods for fast resolution of elliptic problems*, SIAM J. Numer. Anal., 29 (1992), pp. 965–986.

[33] H. Johnen and K. Scherer, *On the equivalence of the K-functional and moduli of continuity and some applications*, in Constructive Theory of Functions of Several Variables, Lecture Notes in Math., Vol. 571, Springer, Berlin, 1977, pp. 119–140.

[34] A. Jouini, *Constructions de Bases d'Ondelettes sur les Variétés*, Dissertation, Université Paris Sud–Centre d'Orsay, Paris, France, 1992.

[35] A. Jouini and P. G. Lemarié-Rieusset, *Analyse multi-résolution bi-orthogonale sur l'intervalle et applications* Ann. Inst. H. Poincaré, Anal. Non Linéaire, 10 (1993), pp. 453–476.

[36] A. Kunoth, *Multilevel preconditioning—appending boundary conditions by Lagrange multipliers*, Adv. Comput. Math., 4 (1995), pp. 145–170.

[37] P. G. Lemarié-Rieusset, *Analyses, multi–résolutions nonorthogonales, commutation entre projecteurs et derivation et ondelettes vecteurs à divergence nulle*, Rev. Mat. Iberoamericana, 8 (1992), pp. 221–237.

[38] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Grundlehren, Vol. 181, Springer-Verlag, New York, Heidelberg, 1973.

[39] Y. Meyer, *Ondelettes et opérateurs* 1–2*: Ondelettes*, Hermann, Paris, 1990.

[40] S. V. Nepomnyaschikh, *Fictitious components and subdomain alternating methods*, Sov. J. Numer. Anal. Math. Model., 5 (1990), pp. 53–68.

[41] P. Oswald, *Multilevel Finite Element Approximations*, Teubner Skripten zur Numerik, Teubner, Stuttgart, 1994.

[42] U. Reif, *TURBS—topologically unrestricted rational B-splines*, Constr. Approx., 14 (1998), pp. 57–77.

[43] T. von Petersdorff, R. Schneider, and C. Schwab, *Multiwavelets for second-kind integral equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2212–2227.

[44] H. Prautzsch, *Freeform splines*, Comput. Aided Geom. Design, 14 (1997), pp. 201–206.

[45] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung großer vollbesetzter Gleichungssysteme*, Habilitationsschrift, Technische Hochschule, Darmstadt, Germany, 1995.

[46] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth-Verlag, Heidelberg, 1995.

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 231–250

STABILITY OF RELATIVE EQUILIBRIA IN THE PROBLEM OF $N+1$ VORTICES*

H. E. CABRAL[†] AND D. S. SCHMIDT[‡]

Abstract. The stability of a regular polygon configuration of N vortices with a central vortex is investigated. When the strength of the central vortex has a value within a certain interval, it is shown that the configuration is locally Liapunov stable. When the stability of the configuration changes, new configurations bifurcate. Although the $N+1$ body problem of celestial mechanics looks similar, it has been shown there that the change of stability and the bifurcation of new configurations occur for different values of the central mass.

Ever since the Adams Prize essay of Thomson, *A Treatise on the Motion of Vortex Rings*, it was known that the stability of the heptagon could not be decided by the linear terms. With methods from fluid mechanics G. J. Mertz had shown in 1978 that the heptagon is stable. By normalizing the Hamiltonian function we can show that except for rotational symmetry the heptagon is locally Liapunov stable.

Key words. vortices, spectral stability, relative equilibria

AMS subject classifications. 76C05, 58F10

PII. S0036141098302124

1. Introduction.

The problem of vortices in an ideal fluid was discussed by J. J. Thomson [13] in his essay for the Adams Prize of 1882. He placed N vortices of equal strength at the vertices of a regular polygon. In a uniformly rotating coordinate system he found that the configuration could be stable for $N \leq 6$ but was unstable for $N \geq 8$. Many papers have extended the work of Thomson since then. Some of the more recent papers include those by G. Morikawa and E. Swenson [9], who studied the motion of geostrophic vortices, and H. Aref [1], who investigated a row of vortices and showed its relationship to the polygon configuration. The question, what happens with a heptagon, is addressed, for example, in [2] and [6].

Maxwell [5] used in his essay for the Adams Prize of 1856 regular polygon configurations for the $N+1$ body problem of celestial mechanics to explain the rings of Saturn. Many papers have been written on that subject, including [7], where bifurcations of relative equilibria in the $N+1$ body problem and $N+1$ vortex problem were studied as a purely algebraic problem without worrying about stability. Moeckel [8] studied carefully the spectral stability of Maxwell's polygon configurations. After this the second author used canonical transformations in [12] in order to explain how bifurcations and changes in stability are related. The current paper is an extension of the above methods to the vortex problem.

By using the methods developed for the problem of celestial mechanics, we are able to derive the results for the polygon configuration of N equal vortices in a straightforward manner. Extending the methods to the polygon configuration with a central vortex of strength κ, that is, the $N+1$ vortex problem, we obtain the results of Thomson and others in a unified way. Actually we can show more, that is, for a

*Received by the editors February 23, 1998; accepted for publication (in revised form) December 17, 1998; published electronically December 21, 1999.

http://www.siam.org/journals/sima/31-2/30212.html

[†] Departamento de Matemática, Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil (hild@dmat.ufpe.br).

[‡] Department of Electrical and Computer Engineering and Computer Science, University of Cincinnati, Cincinnati, OH 45221-0030 (dieter.schmidt@uc.edu).

bounded range of κ the $N+1$ vortex configuration is not only linearly stable, but it is also locally stable in the sense of Liapunov. For seven bodies this interval of stability starts when $\kappa = 0$. This explains the special nature of the heptagon. By normalizing the Hamiltonian function through fourth-order terms we show that the heptagon of Thomson is also locally Liapunov stable.

Since the $N+1$ vortex problem has only $N+1$ degrees of freedom versus $2(N+1)$ for the problem in celestial mechanics, the calculations are simpler. Furthermore, polygon configuration have a lot of symmetries, which make these calculations feasible. In addition to this a Taylor series expansion for the logarithmic potential of the vortex problem can be found with the help of analytic functions, which makes it much easier to compute than the one for the Newtonian potential.

It is for this reason that we have repeated the calculations of [12] and [7], not only to work on a different problem but with the hope that the results of this paper can be used later to get some additional insight into the corresponding problem of celestial mechanics.

We consider the planar flow of an ideal incompressible fluid, which is not constrained by any boundaries. For any positively oriented, closed path, circulation is defined by

$$\Gamma = \oint v_s ds,$$

where v_s represents the component of velocity along an element of the path of length ds. The strength of the vortices enclosed by this path is then given by

$$\kappa = \frac{\oint v_s ds}{Area}.$$

Let $N+1$ vortices be located at $q_j = (x_j, y_j)$ with vortex strength κ_j for $j = 0, \dots, N$. Kirchhoff [3] derived the equations of motion for these vortices in the framework of Hamiltonian mechanics and these equations are

$$\kappa_j \dot{x}_j = \frac{\partial U}{\partial y_j},$$
$$\kappa_j \dot{y}_j = -\frac{\partial U}{\partial x_j},$$

with the Hamiltonian function

$$U = -\sum_{i<j} \kappa_i \kappa_j \log\left((x_i - x_j)^2 + (y_i - y_j)^2\right)^{\frac{1}{2}}. \tag{1.1}$$

The potential function of celestial mechanics has a similar form. Therefore, we call U the logarithmic potential function and hope that it will not be confused with the velocity potential in the theory of incompressible potential flow.

With the column vector $q = (x_0, \dots, x_N, y_0, \dots, y_N)^T$, the diagonal matrix $\mathbf{M} = \mathrm{diag}(\kappa_0, \dots, \kappa_N, \kappa_0, \dots, \kappa_N)$, and the standard symplectic matrix

$$\mathbf{J} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix},$$

the above equation can also be written in vector form

$$\mathbf{M}\dot{q} = \mathbf{J}\,\nabla U. \tag{1.2}$$

A relative equilibrium is a configuration of vortices, which becomes stationary in a rotating coordinate system. Let $Q \in \mathcal{R}^{2(N+1)}$ be the coordinates, which rotate uniformly with angular velocity ν around the origin, so that the coordinate transformation is given by

$$q = \mathrm{e}^{\nu \mathbf{J} t} Q.$$

The equations of motion are then

$$\mathbf{M}\dot{Q} = \mathbf{J}(-\nu \mathbf{M} Q + \nabla U(Q)). \tag{1.3}$$

A stationary solution Q_0 satisfies

$$0 = -\nu \mathbf{M} Q_0 + \nabla U(Q_0).$$

Form the scalar product of the above equation with Q_0 to find for the value of ν

$$\nu = \frac{Q_0^T \nabla U(Q_0)}{Q_0^T \mathbf{M} Q_0}. \tag{1.4}$$

Since

$$U(\lambda Q) = -\log \lambda \sum \kappa_i \kappa_j + U(Q), \qquad \lambda > 0,$$

differentiating with respect to λ and setting $\lambda = 1$, we get

$$Q^T \cdot \nabla U(Q) = -\sum_{i<j} \kappa_i \kappa_j,$$

which lets us find a simpler expression for the numerator in (1.4). For the $N+1$ body problem of celestial mechanics, the denominator of (1.4) would be related to the moment of inertia. We will use the same notation and write

$$I(Q_0) = \frac{1}{2} Q_0^T \mathbf{M} Q_0 = \frac{1}{2} \sum_{j=0}^{N} \kappa_j (x_{0,j}^2 + y_{0,j}^2).$$

The angular velocity for the uniformly rotating coordinate system is then

$$\nu = \frac{-\sum_{i<j} \kappa_i \kappa_j}{2I(Q_0)}.$$

The derivation above demonstrates the similarity of the problem of $N+1$ vortices with the $N+1$ problem of celestial mechanics. Of these two problems the vortex problem is actually simpler since the position coordinates x_j and y_j are already conjugate to each other. In a strict sense (1.2) is not a Hamiltonian differential equation; nevertheless, most results in Hamiltonian mechanics can be carried over, as one can consider a nonstandard symplectic structure (see [4] or [10]). Thus it is not surprising that the computations for the spectral stability of relative equilibria in the $N+1$ body problem (see [8], [12]) can also be repeated for the vortex problem. What is surprising is that the results are more degenerate than those in celestial mechanics.

In the next section we give the conditions for spectral stability of relative equilibrium and mention the problems of a ring of vortices and the ring with a central vortex. These two problems will be discussed in the following sections. We then vary the strength κ of the central vortex and show in the last section that the change of stability occurs at those values of κ where new relative equilibria bifurcate. The final section reports on our calculations when $N = 7$; that is, we demonstrate that the heptagon configuration is locally Liapunov stable.

2. Spectral stability. In order to determine the stability of the relative equilibrium Q_0 of (1.3) we look at solutions nearby and set

$$Q = Q_0 + Z,$$

so that (1.3) becomes

$$\mathbf{M}\dot{Z} = \mathbf{J}\nabla V(Z), \tag{2.1}$$

where

$$V(Z) = -\frac{\nu}{2}(Q_0 + Z)^T\mathbf{M}(Q_0 + Z) + U(Q_0 + Z). \tag{2.2}$$

Since Q_0 is a stationary point, $\nabla V(0) = -\nu\mathbf{M}Q_0 + \nabla U(Q_0) = 0$. With $D^2V(0) = -\nu\mathbf{M} + D^2U_0$, where D^2U_0 stands for the Hessian of U evaluated at Q_0, the linearized form of (2.1) is

$$\mathbf{M}\dot{Z} = \mathbf{J}(-\nu\mathbf{M} + D^2U_0)Z. \tag{2.3}$$

Solutions of the form $Z = \mathrm{e}^{\lambda t}\zeta$ exist when the matrix

$$\lambda\mathbf{J} - \nu\mathbf{I} + \mathbf{M}^{-1}D^2U_0 \tag{2.4}$$

is singular.

Definition 2.1. *A relative equilibrium solution Q_0 of* (1.3) *is spectrally stable if all roots λ of*

$$\left|\lambda\mathbf{J} - \nu\mathbf{I} + \mathbf{M}^{-1}D^2U_0\right| = 0 \tag{2.5}$$

are purely imaginary.

Remark. In order for the relative equilibrium to be linearly stable, we require, in addition, that all elementary divisors have simple roots; that is, the system (2.3) can be diagonalized completely.

The problem to be considered is a ring of N identical vortices at the vertices of a regular N-gon with an additional vortex at the origin. Let $\omega = \mathrm{e}^{2\pi\sqrt{-1}/N}$ be the Nth root of unity. The position of the jth vortex is then ω^j for $j = 0, \ldots, N-1$ and its strength is κ_0. The strength of the vortex at the origin will be κ_N. We write therefore for the potential function (1.1)

$$U = \kappa_0^2 U_1 + \kappa_0\kappa_N U_2,$$

where the two functions are given by

$$U_1 = -\sum_{0\le i<j<N} \log|Q_i - Q_j| = -\frac{1}{2}\sum_{i=0}^{N-1}\sum_{j=1}^{N-1} \log|Q_i - Q_{i+j}| \tag{2.6}$$

with the index $i + j$ to be taken mod$N$ and

$$U_2 = -\sum_{j=0}^{N-1} \log|Q_j - Q_N|. \tag{2.7}$$

If in (2.4) we scale λ and ν by κ_0, then only the ratio of the strength of the two vortices κ_N/κ_0 is of significance. Thus without loss of generality we can set $\kappa_0 = 1$

and write κ instead of κ_N for the strength of the vortex at the origin. The potential function to be considered is therefore

$$U = U_1 + \kappa U_2. \tag{2.8}$$

The moment of inertia for the above configuration is then $I(Q_0) = N/2$, and its rate of rotation is

$$\nu = -\frac{N-1}{2} - \kappa. \tag{2.9}$$

3. Expansion of the Hamiltonian function U_1. Consider first the case of a ring where the N vortices form a regular N-gon. Since $\kappa = 0$, we can ignore here the $(N+1)$st coordinate and consider the problem in $\mathcal{R}^{2N}$. We will use a transformation which brings $D^2U_1(Q_0)$ into a normal form. The transformation has been used in a nonsymplectic form by Palmore [11] and others. In complex notation the transformation is $Z = \mathbf{W}z$ or

$$Q = Q_0 + \mathbf{W}z \tag{3.1}$$

with

$$\mathbf{W} = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{N-1} \\ \vdots & \vdots & & \vdots \\ 1 & \omega^{N-1} & \cdots & \omega^{(N-1)^2} \end{pmatrix}. \tag{3.2}$$

The matrix $\mathbf{W}$ is unitary and therefore defines a symplectic transformation in $\mathcal{R}^{2N}$. This is seen when we keep the correspondence of the complex-valued ω with the 2×2 real submatrices

$$\begin{pmatrix} \cos 2\pi/N & -\sin 2\pi/N \\ \sin 2\pi/N & \cos 2\pi/N \end{pmatrix}$$

in mind. Since $\mathbf{M}$ is an identity matrix, the transformed differential equation (2.3) is

$$\dot{z} = (-\nu\mathbf{J} + \mathbf{J}\mathbf{W}^T D^2 U_1 \mathbf{W})z,$$

where $\mathbf{W}^T$ stands for the transpose of the $2N \times 2N$ matrix $\mathbf{W}$ written in its real form. The only task, albeit a tedious one, is the transformation of the Hessian. We accomplish it by computing partial derivatives of $U_1(Q_0 + \mathbf{W}z)$ and evaluating them at $z = 0$. This job is simplified by staying with complex notation as long as possible.

Consider first a function $f(z) = \log \phi(z)$ of the complex variable $z = x + \sqrt{-1}y$. The partial derivatives of the function $g(x, y) = \log|\phi|$ can be obtained from the real part of the derivative with respect to z of the analytical function $f(z)$ since

$$\log \phi = \log|\phi| + \sqrt{-1}\arg\phi. \tag{3.3}$$

Therefore, write $z_k = x_k + \sqrt{-1}y_k$ for $k = 0, 1, \ldots, N-1$. From (3.2) we have

$$Q_j = \omega^j + \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \omega^{jk} z_k$$

and we can write

$$\begin{aligned} Q_i - Q_{i+j} &= \omega^i - \omega^{i+j} + \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} (\omega^{ik} - \omega^{(i+j)k}) z_k \\ &= \omega^i \left(d_j + \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} d_{jk} \omega^{i(k-1)} z_k \right), \end{aligned}$$

where we have set

$$d_j = 1 - \omega^j \tag{3.4}$$

so that $|d_j| = 2 \sin \pi j/N$. Since $|\omega| = 1$ for ϕ in (3.3) use

$$\phi = (Q_i - Q_{i+j})\omega^{-i} = d_j + \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} d_{jk} \omega^{i(k-1)} z_k$$

so that

$$\frac{\partial \phi}{\partial z_k} = \frac{d_{jk}}{\sqrt{N}} \omega^{i(k-1)}. \tag{3.5}$$

For $f = \log \phi$, we then find

$$\begin{aligned} \frac{\partial f}{\partial z_r} &= \frac{1}{\phi} \frac{\partial \phi}{\partial z_r}, \\ \frac{\partial^2 f}{\partial z_r \partial z_s} &= \frac{-1}{\phi^2} \frac{\partial \phi}{\partial z_r} \frac{\partial \phi}{\partial z_s}, \\ \frac{\partial^3 f}{\partial z_r \partial z_s \partial z_t} &= \frac{2}{\phi^3} \frac{\partial \phi}{\partial z_r} \frac{\partial \phi}{\partial z_s} \frac{\partial \phi}{\partial z_t}, \end{aligned}$$

and so on. At $z = 0$ we have $\phi(0) = d_j$ and the partial derivatives are therefore

$$\begin{aligned} \frac{\partial f}{\partial z_r} &= \frac{1}{\sqrt{N}} \omega^{i(r-1)} \frac{d_{jr}}{d_j}, \\ \frac{\partial^2 f}{\partial z_r \partial z_s} &= \frac{-1}{N} \omega^{i(r+s-2)} \frac{d_{jr} d_{js}}{d_j^2}, \\ \frac{\partial^3 f}{\partial z_r \partial z_s \partial z_t} &= \frac{2}{N\sqrt{N}} \omega^{i(r+s+t-3)} \frac{d_{jr} d_{js} d_{jt}}{d_j^3}, \\ \frac{\partial^4 f}{\partial z_r \partial z_s \partial z_t \partial z_u} &= \frac{-6}{N^2} \omega^{i(r+s+t+u-4)} \frac{d_{jr} d_{js} d_{jt} d_{ju}}{d_j^4}, \end{aligned}$$

and so on. We then obtain, for example, the second-order partial derivatives of $g = \log |\phi|$ to be

$$\frac{\partial^2 g}{\partial x_r \partial x_s} = -\frac{1}{N} \mathrm{Re} \left(\omega^{i(r+s-2)} \frac{d_{jr} d_{js}}{d_j^2} \right).$$

Similarly, we find

$$\frac{\partial^2 g}{\partial y_r \partial x_s} = \frac{1}{N} \mathrm{Im} \left(\omega^{i(r+s-2)} \frac{d_{jr} d_{js}}{d_j^2} \right) \quad \text{and} \quad \frac{\partial^2 g}{\partial y_r \partial y_s} = -\frac{\partial^2 g}{\partial x_r \partial x_s}.$$

Since ω is the Nth root of unity,

$$\sum_{i=0}^{N-1} \omega^{i(r+s-2)} = \begin{cases} N & \text{for} \quad r+s=2 \bmod N, \\ 0 & \text{otherwise,} \end{cases}$$

so that

$$\frac{\partial^2 U_1}{\partial x_r \partial x_s} = -\frac{1}{2}\sum_{i=0}^{N-1}\sum_{j=1}^{N-1} \frac{\partial^2 g}{\partial x_r \partial x_s} = \frac{1}{2}\sum_{j=1}^{N-1} \mathrm{Re}\left(\frac{d_{jr} d_{js}}{d_j^2}\right)$$

for $r+s=2 \bmod N$ and zero otherwise. On the other hand,

$$\sum_{j=1}^{N-1} \frac{d_{jr} d_{js}}{d_j^2} = \sum_{j=1}^{N-1} \frac{(1-\omega^{jr})(1-\omega^{js})}{(1-\omega^j)^2};$$

hence, this sum is zero if $r=0$ or $s=0$, while in the other cases it is equal to

$$\sum_{j=1}^{N-1} (1+\omega^j+\cdots+\omega^{j(r-1)})(1+\omega^j+\cdots+\omega^{j(s-1)}).$$

Since

$$\sum_{j=1}^{N-1} \omega^{jm} = -1 \quad \text{if} \quad m \neq 0 \bmod N,$$

and denoting by $k(r,s)$ the number of pairs (ρ,σ) such that $\rho+\sigma=0 \bmod N$, with $0 \le \rho < r$ and $0 \le \sigma < s$, we see that this sum is equal to

$$k(r,s)(N-1)-(rs-k(r,s)) = N\ k(r,s)-rs.$$

We readily see that there are only two cases, one being $r=s=1$ with $k(1,1)=1$ and the other $r+s=N+2$ when r or s is greater than 1, which gives $k(r,s)=2$. Thus,

$$\frac{\partial^2 U_1}{\partial x_1^2} = -\frac{\partial^2 U_1}{\partial y_1^2} = \frac{N-1}{2}, \tag{3.6}$$

and, for $r+s=2 \bmod N$, $rs \neq 0$,

$$\frac{\partial^2 U_1}{\partial x_r \partial x_s} = -\frac{\partial^2 U_1}{\partial y_r \partial y_s} = \frac{2N-rs}{2}, \tag{3.7}$$

while the derivatives are zero in all the other cases.

By summing up the terms for the third-order partial derivatives we will see, as before, that most of them cancel out and we are left, for $r+s+t=3 \bmod N$, with the expression

$$\frac{\partial^3 U_1}{\partial x_r \partial x_s \partial x_t} = \frac{-1}{\sqrt{N}}\mathrm{Re}\left(\sum_{j=1}^{N-1} \frac{d_{jr} d_{js} d_{jt}}{d_j^3}\right),$$

and this derivative is zero if $r+s+t \neq 3 \bmod N$ or if $rst = 0$. Similarly, we find that

$$\frac{\partial^3 U_1}{\partial y_r \partial y_s \partial x_t} = -\frac{\partial^3 U_1}{\partial x_r \partial x_s \partial x_t},$$

while the third-order derivatives involving one or three y's are zero.

Denoting by $k(r,s,t)$ the number of triplets (ρ, σ, τ) such that $\rho+\sigma+\tau = 0 \bmod N$ when $0 \leq \rho < r$, $0 \leq \sigma < s$, $0 \leq \tau < t$, we find that

$$\frac{\partial^3 U_1}{\partial x_r \partial x_s \partial x_t} = -\frac{1}{\sqrt{N}}(N\ k(r,s,t) - rst).$$

We find similar expressions for the fourth-order partial derivatives of U_1, for instance,

$$\frac{\partial^4 U_1}{\partial x_r \partial x_s \partial x_t \partial x_u} = \frac{3}{N}\mathrm{Re}\left(\sum_{j=1}^{N-1} \frac{d_{jr} d_{js} d_{jt} d_{ju}}{d_j^4}\right) \quad \text{if} \quad r+s+t+u = 4 \bmod N.$$

We compute

$$\frac{\partial^4 U_1}{\partial x_r \partial x_s \partial x_t \partial x_u} = -\frac{\partial^4 U_1}{\partial x_r \partial x_s \partial y_t \partial y_u} = \frac{\partial^4 U_1}{\partial y_r \partial y_s \partial y_t \partial y_u} = \frac{3}{N}(N\ k(r,s,t,u) - rstu)$$

if $r+s+t+u = 4 \bmod N$, $rstu \neq 0$. The remaining partial derivatives are all equal to zero.

4. Stability polynomial for a ring of vortices. Let us call the nonzero second-order derivatives of U_1 in (3.7) $c_r = \frac{1}{2}(2N - rs)$, which holds when $r+s = 2+N$ so that

$$c_r = -\frac{1}{2}(r-2)(N-r) \tag{4.1}$$

for $r = 2, 3, \ldots$. For $r = 1$, the above formula gives also the correct value of (3.6) so that (4.1) is valid for all cases. For reference we list in Table 4.1 the values of c_r for 3 to 12 vortices. Actually only about half of the values are listed, as the remaining ones are found via symmetry from $c_r = c_{N+2-r}$.

Table 4.1
Values of c_r given by (4.1) *for N vortices.*

N	c_1	c_2	c_3	c_4	c_5	c_6	c_7
3	1.0	0.0					
4	1.5	0.0	−0.5				
5	2.0	0.0	−1.0				
6	2.5	0.0	−1.5	−2.0			
7	3.0	0.0	−2.0	−3.0			
8	3.5	0.0	−2.5	−4.0	−4.5		
9	4.0	0.0	−3.0	−5.0	−6.0		
10	4.5	0.0	−3.5	−6.0	−7.5	−8.0	
11	5.0	0.0	−4.0	−7.0	−9.0	−10.0	
12	5.5	0.0	−4.5	−8.0	−10.5	−12.0	−12.5

Hence we have the matrix

$$\mathbf{C} = \left(\frac{\partial^2 U_1}{\partial x_r \partial x_s}\right) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & c_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & c_3 \\ 0 & 0 & 0 & 0 & 0 & \cdots & c_4 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & c_4 & \cdots & 0 & 0 \\ 0 & 0 & 0 & c_3 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

The matrix (2.4) has thus the block form

$$\begin{pmatrix} -\nu\mathbf{I} + \mathbf{C} & \lambda\mathbf{I} \\ -\lambda\mathbf{I} & -\nu\mathbf{I} - \mathbf{C} \end{pmatrix} \tag{4.2}$$

with $\mathbf{I}$ an $N \times N$ identity matrix. Due to the special nature of $\mathbf{C}$ the determinant of (4.2) decomposes into the product of 2×2 and 4×4 subdeterminants, which is easily seen by shuffling rows and columns. These cases will be referred to by the index r of the coefficient in the matrix $\mathbf{C}$ and discussed now as follows.

Cases $r = 0$ and $r = 2$.

$$\begin{vmatrix} -\nu & \lambda \\ -\lambda & -\nu \end{vmatrix} = \lambda^2 + \nu^2 = 0.$$

It gives rise to two pairs of trivial solutions $\lambda = \pm \mathrm{i}\nu$, which have to be expected due to the existence of four integrals.

Case $r = 1$.

$$\begin{vmatrix} -\nu + c_1 & \lambda \\ -\lambda & -\nu - c_1 \end{vmatrix} = \lambda^2 + \nu^2 - c_1^2 = \lambda^2 = 0.$$

This pair of zero roots is related to the fact that the relative equilibria are circular solutions in an inertial coordinate system and that among all solutions the circular ones are isolated. The remaining cases give the nontrivial ones.

Cases $r > 2$, $s = N + 2 - r$, and $c_r = c_s$.

$$\begin{vmatrix} -\nu & c_r & \lambda & 0 \\ c_r & -\nu & 0 & \lambda \\ -\lambda & 0 & -\nu & -c_r \\ 0 & -\lambda & -c_r & -\nu \end{vmatrix} = (\lambda^2 + \nu^2 - c_r^2)^2 = 0$$

has the real double roots $\lambda = \pm\sqrt{c_r^2 - \nu^2}$. At these values the corresponding 4×4 matrix has rank two, and therefore the system can be diagonalized completely. Finally when N is even, the following determinant has to be considered.

Case $r = s = \frac{N+2}{2}$.

$$\begin{vmatrix} -\nu + c_r & \lambda \\ -\lambda & -\nu - c_r \end{vmatrix} = \lambda^2 + \nu^2 - c_r^2 = 0,$$

which gives the simple pair of real roots $\lambda = \pm\sqrt{c_r^2 - \nu^2}$.

Since $c_r^2 - \nu^2 = c_r^2 - c_1^2 = \{(r-2)(N-r) + (N-1)\}\{(r-2)(N-r) - (N-1)\}/4$, we determine from this formula or by inspecting Table 4.1 the number of different characteristic exponents. Table 4.2 gives these numbers with the trivial characteristic exponents included in the count. As already discovered by Thomson the ring with 7 vortices is somewhat degenerate, and we will get back to this case later on.

Table 4.2
Number of roots of the stability polynomial for a ring with N vortices.

N	Negative	Positive	Zero	Imaginary
3	0	0	2	4
4	0	0	2	6
5	0	0	2	8
6	0	0	2	10
7	0	0	6	8
≥ 8	$N-5$	$N-5$	2	8

5. Stability of a ring with a central vortex. When we include a central vortex we have to use the function (2.8) and have to work in $\mathcal{R}^{2N+2}$. We can use the results of the previous section but have to add to it the contributions of the potential function U_2 given in (2.7). As an intermediate step we first determine for $0 \leq j \leq N-1$ the partial derivatives of $f = \log \phi$ at $z = 0$, where

$$\phi = (Q_j - Q_N)\omega^{-j} = 1 + \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \omega^{j(k-1)} z_k - \omega^{-j} z_N$$

with the new coordinate $Q_N = z_N$. The derivatives of first order are

$$\frac{\partial f}{\partial z_r} = \begin{cases} \dfrac{1}{\sqrt{N}} \omega^{j(r-1)}, & 0 \leq r \leq N-1, \\ -\omega^{-j}, & r = N, \end{cases}$$

and of the second order

$$\frac{\partial^2 f}{\partial z_r \partial z_s} = \begin{cases} -\dfrac{1}{N} \omega^{j(r+s-2)}, & 0 \leq r, s \leq N-1, \\ \dfrac{1}{\sqrt{N}} \omega^{j(r-2)}, & 0 \leq r \leq N-1, \quad s = N, \\ -\omega^{-2j}, & r = s = N. \end{cases}$$

By summing over j from 0 through $N-1$, we have that the only nonzero terms left are

$$\frac{\partial^2 U_2}{\partial x_r \partial x_s} = \begin{cases} 1 & \text{for } r + s = 2 \bmod N, \\ -\sqrt{N} & \text{for } r = 2 \text{ and } s = N, \text{ or } r = N \text{ and } s = 2, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix (2.4) has again the form given in (4.2) with $\mathbf{I}$ now an $(N+1) \times (N+1)$ identity matrix, $\nu = -c_1 - \kappa$, and after the multiplication with $\mathbf{M}^{-1}$ the matrix $\mathbf{C}$ is replaced by

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & \kappa & 0 & \cdots & 0 & 0 \\ 0 & c_1 + \kappa & 0 & 0 & \cdots & 0 & 0 \\ \kappa & 0 & 0 & 0 & \cdots & 0 & -\kappa\sqrt{N} \\ 0 & 0 & 0 & 0 & \cdots & c_3 + \kappa & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & c_3 + \kappa & \cdots & 0 & 0 \\ 0 & 0 & -\sqrt{N} & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

The splitting of (4.2) into subdeterminants still occurs. For rows $r = 0, 2$, and N we obtain the following subdeterminant:

$$\begin{vmatrix} -\nu & \kappa & 0 & \lambda & 0 & 0 \\ \kappa & -\nu & -\kappa\sqrt{N} & 0 & \lambda & 0 \\ 0 & -\sqrt{N} & -\nu & 0 & 0 & \lambda \\ -\lambda & 0 & 0 & -\nu & -\kappa & 0 \\ 0 & -\lambda & 0 & -\kappa & -\nu & \kappa\sqrt{N} \\ 0 & 0 & -\lambda & 0 & \sqrt{N} & -\nu \end{vmatrix} = 0.$$

For block matrices the following is easily verified:

$$\begin{vmatrix} \mathbf{A} & \lambda\mathbf{I} \\ -\lambda\mathbf{I} & \mathbf{B} \end{vmatrix} = \left|\lambda^2\mathbf{I} + \mathbf{AB}\right|,$$

so that for the determinant above we find

$$\begin{vmatrix} \lambda^2+\nu^2-\kappa^2 & 0 & \kappa^2\sqrt{N} \\ 0 & \lambda^2+\nu^2-\kappa^2-\kappa N & 0 \\ \kappa\sqrt{N} & 0 & \lambda^2+\nu^2-\kappa N \end{vmatrix}$$

$$= (\lambda^2+\nu^2)(\lambda^2+\nu^2-\kappa^2-\kappa N)^2 = 0.$$

The first factor gives the trivial exponents, whereas the second factor gives the repeated characteristic exponents

$$\lambda = \pm\sqrt{\kappa - (N-1)^2/4}.$$

Thus a change in stability occurs when

$$\kappa = \left(\frac{N-1}{2}\right)^2, \tag{5.1}$$

with the configuration becoming less stable as κ increases. The case of $r = 1$ leads to a determinant of the form

$$\begin{vmatrix} -2\nu & \lambda \\ -\lambda & 0 \end{vmatrix} = 0,$$

so that it again gives the repeated zeros. The general case $r > 2$ gives with $\nu = -c_1 - \kappa$ the 4×4 submatrix

$$\begin{pmatrix} c_1+\kappa & c_r+\kappa & \lambda & 0 \\ c_r+\kappa & c_1+\kappa & 0 & \lambda \\ -\lambda & 0 & c_1+\kappa & -c_r-\kappa \\ 0 & -\lambda & -c_r-\kappa & c_1+\kappa \end{pmatrix} \tag{5.2}$$

whose determinant could be evaluated as above. Instead we use an additional symplectic transformation; that is, we perform the rotation

$$\begin{pmatrix} x_r \\ x_s \end{pmatrix} \leftarrow \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_r \\ x_s \end{pmatrix}, \qquad \begin{pmatrix} y_r \\ y_s \end{pmatrix} \leftarrow \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} y_r \\ y_s \end{pmatrix}.$$

The matrix (5.2) then becomes and reads

$$\begin{pmatrix} c_1 - c_r & 0 & \lambda & 0 \\ 0 & c_1 + c_r + 2\kappa & 0 & \lambda \\ -\lambda & 0 & c_1 + c_r + 2\kappa & 0 \\ 0 & -\lambda & 0 & c_1 - c_r \end{pmatrix}$$

from where we obtain the repeated characteristic exponents

$$\lambda = \pm\sqrt{(c_r - c_1)(2\kappa + c_r + c_1)}. \tag{5.3}$$

A change of stability occurs when

$$\kappa = -\frac{c_r + c_1}{2} \tag{5.4}$$

with the configuration becoming more stable as κ increases. Finally we consider the case when N is even and a 2×2 subdeterminant occurs for $r = s = (N+2)/2$:

$$\begin{vmatrix} c_1 + c_r + 2\kappa & \lambda \\ -\lambda & c_1 - c_r \end{vmatrix} = 0.$$

It gives rise to characteristic exponents that have the same form as in (5.3), except that they are now simple.

A change of stability occurs at values of κ given in (5.1) and (5.4). Since these values do not depend only on r but also on N, we denote them by

$$\kappa(r, N) = \begin{cases} -\dfrac{c_r + c_1}{2} & \text{for } 2 < r \le (N+2)/2, \\ \left(\dfrac{N-1}{2}\right)^2 & \text{for } r = 2. \end{cases} \tag{5.5}$$

The ordering of the values where the stability of the $N+1$ vortex configuration changes is

$$-0.5 = \kappa(3, N) < \kappa(4, N) < \cdots < \kappa\left(\left\lfloor \frac{N+2}{2} \right\rfloor, N\right) < \kappa(2, N) = (N-1)^2/4.$$

Values for the $\kappa(r, N)$ are listed in this order in Table 5.1. Also listed under $\kappa_{v=0}$ is the value of κ, where the rate of rotation ν of the coordinate system is zero.

Table 5.1
Critical values for $\kappa(r, N)$ for a ring with a central vortex.

N	$\kappa_{\nu=0}$	$\kappa(3,N)$	$\kappa(4,N)$	$\kappa(5,N)$	$\kappa(6,N)$	$\kappa(7,N)$	$\kappa(2,N)$
3	−1.0						1.00
4	−1.5	−0.5					2.25
5	−2.0	−0.5					4.00
6	−2.5	−0.5	−0.25				6.25
7	−3.0	−0.5	0.00				9.00
8	−3.5	−0.5	0.25	0.50			12.25
9	−4.0	−0.5	0.50	1.00			16.00
10	−4.5	−0.5	0.75	1.50	1.75		20.25
11	−5.0	−0.5	1.00	2.00	2.50		25.00
12	−5.5	−0.5	1.25	2.50	3.25	3.50	30.25

Furthermore, (5.3) shows that the configuration becomes more stable as κ passes through one of the values $\kappa(r, N)$ for $r > 2$, since we have $c_r - c_1 < 0$. On the other hand (5.1) shows that we lose stability when $\kappa > \kappa(2, N)$. This contradicts the claim made in [6] that a very strong central vortex could make the configuration stable. This happens in the $N + 1$ body problem of celestial mechanics with a large mass of the central body when $N > 7$. For the $N + 1$ vortex configuration, spectral stability can occur only for values for the central vortex in the interval $\kappa(\lfloor (N + 2)/2 \rfloor, N) \leq \kappa \leq \kappa(2, N)$. We can prove even more in Theorem 5.1.

THEOREM 5.1. *The $N + 1$ vortex configuration with the central vortex of size κ is Liapunov stable for*
$(N^2 - 8N + 8)/16 < \kappa < (N - 1)^2/4$ *when N is even, and*
$(N^2 - 8N + 7)/16 < \kappa < (N - 1)^2/4$ *when N is odd.*

Proof. The values on the left-hand side of the above inequalities are those of $\kappa(\lfloor (N+2)/2 \rfloor)$ and on the right-hand side those of $\kappa(2, N)$. As was shown above only in this interval are all nontrivial roots of the characteristic polynomial purely imaginary. We will now show that in this interval the quadratic part of the Hamiltonian is positive definite and can serve as a Liapunov function.

After a sequence of symplectic transformations, including the rotation mentioned above, the original system of equations (1.2) was transformed to represent the motion near the equilibrium by

$$\mathbf{M}\dot{z} = \mathbf{J}\nabla H(z)$$

with the column vector $z = (x, y)^T$ in $\mathcal{R}^{2(N+1)}$. The Hamiltonian $H(z)$ starts with the second-order terms:

$$H_2(z) = \frac{1}{2}(x^T \mathbf{A} x + y^T \mathbf{B} y).$$

Since the matrices $\mathbf{A}$ and $\mathbf{B}$ are mostly diagonal, we will not display them but write down the Hamiltonian with the components as they were used in the discussion of the different cases

$$\begin{aligned} H_2(z) = \frac{1}{2}\,[x_0, x_2, x_N] & \begin{bmatrix} c_1 + \kappa & \kappa & 0 \\ \kappa & c_1 + \kappa & -\kappa\sqrt{N} \\ 0 & -\sqrt{N} & c_1 + \kappa \end{bmatrix} \begin{bmatrix} x_0 \\ x_2 \\ x_N \end{bmatrix} + (c_1 + \kappa)x_1^2 \\ & + \frac{c_1 + c_3 + 2\kappa}{2} x_3^2 + \frac{c_1 + c_4 + 2\kappa}{2} x_4^2 + \cdots + \frac{c_1 - c_4}{2} x_{N-2}^2 + \frac{c_1 - c_3}{2} x_{N-1}^2 \\ & + \frac{1}{2}\,[y_0, y_2, y_N] \begin{bmatrix} c_1 + \kappa & -\kappa & 0 \\ -\kappa & c_1 + \kappa & \kappa\sqrt{N} \\ 0 & \sqrt{N} & c_1 + \kappa \end{bmatrix} \begin{bmatrix} y_0 \\ y_2 \\ y_N \end{bmatrix} \\ & + \frac{c_1 - c_3}{2} y_3^2 + \frac{c_1 - c_4}{2} y_4^2 + \cdots + \frac{c_1 + c_4 + 2\kappa}{2} y_{N-2}^2 + \frac{c_1 + c_3 + 2\kappa}{2} y_{N-1}^2. \end{aligned}$$

In the given interval for κ all coefficients of the quadratic terms are positive. It remains to show that the two quadratic forms are also positive definite. The reason why these matrices are not diagonalized and left as they are has to do with the fact that a transformation matrix has also to commute with $\mathbf{M}$. Nevertheless, their three eigenvalues are easily computed and they are

$$c_1 + \kappa, \qquad c_1 + \kappa + \sqrt{\kappa(\kappa + N)}, \qquad c_1 + \kappa - \sqrt{\kappa(\kappa + N)}. \tag{5.6}$$

For $0 < \kappa < (N-1)^2$ only the sign of the last eigenvalue is questionable. With $c_1 = (N-1)/2$ consider therefore the function

$$\lambda(\kappa) = (N-1)/2 + \kappa - \sqrt{\kappa(\kappa+N)}.$$

With

$$\lambda'(\kappa) = 1 - \frac{2\kappa + N}{2\sqrt{\kappa(\kappa+N)}} < 0 \qquad \text{for} \quad \kappa > 0$$

and $\lambda(0) > 0$ we look for a positive root of $\lambda(\kappa) = 0$ by squaring

$$(N-1)/2 + \kappa = \sqrt{\kappa(\kappa+N)}.$$

We find this value to be $\kappa = (N-1)^2/4 = c_1^2$, which coincides with the upper limit of the interval under consideration.

For $N < 7$ the given intervals for κ allow negative values for the central vortex. These intervals are

$$\begin{aligned}
N &= 3, \quad -1.0 < \kappa < 1.00, \quad c_1 = 1.0,\\
N &= 4, \quad -0.5 < \kappa < 2.25, \quad c_1 = 1.5,\\
N &= 5, \quad -0.5 < \kappa < 4.00, \quad c_1 = 2.0,\\
N &= 6, \quad -0.25 < \kappa < 6.25, \quad c_1 = 2.5,
\end{aligned}$$

with the value of c_1 listed in each case. Since the real part of the eigenvalues (5.6) are positive, it follows that the quadratic forms are positive definite.

6. Degenerate relative equilibria. Relative equilibria are characterized as the extrema of the potential function U under the condition that the moment of inertia is kept constant. In [7] it was shown that degenerate relative equilibria lead to new configurations. Since the discussion in that paper is somewhat terse and its notation different, we will repeat the arguments here.

A configuration is called degenerate if it is not isolated in a reduced space. The configurations discussed so far are not isolated in a trivial manner, since we can move the center of vorticity by a finite amount and find another one. We could also rotate the configuration around its center by a finite angle. In order to remove these trivial degeneracies one has to work in a quotient space that removes these actions. Keeping the center of vorticity at the origin gives in our coordinates

$$z_0 = -\frac{\kappa}{\sqrt{N}} z_N.$$

The rotation is removed by requiring that z_1 is real; i.e., $y_1 = 0$. The moment of inertia of the unperturbed N-gon is $N/2$, and keeping it fixed leads to the equation

$$0 = 2\sqrt{N} x_1 + x_1^2 + \sum_{k=2}^{N-1} (x_k^2 + y_k^2) + \frac{\kappa}{N}(N+\kappa)(x_N^2 + y_N^2).$$

By the implicit function theorem this equation can be solved to give

$$x_1 = x_1(x_2, \ldots, x_N, y_2, \ldots, y_N)$$

near the origin. All of its first-order and most of its second-order partial derivatives are zero with the exception of

$$\frac{\partial^2 x_1}{\partial x_r^2} = \frac{\partial^2 x_1}{\partial y_r^2} = -\frac{1}{\sqrt{N}} \qquad \text{for} \quad 2 \le r \le N-1$$

and

$$\frac{\partial^2 x_1}{\partial x_N^2} = \frac{\partial^2 x_1}{\partial y_N^2} = -\frac{\kappa}{\sqrt{N}}\left(1 + \frac{\kappa}{N}\right).$$

If $\tilde{U}$ is the restriction of U given by these formulas, that is,

$$\tilde{U}(x_2, \dots, y_N) = U\left(-\frac{\kappa x_N}{\sqrt{N}}, x_1, x_2, \dots, x_N, -\frac{\kappa y_N}{\sqrt{N}}, 0, y_2, \dots, y_N\right),$$

then to compute its Hessian at $z = 0$, we also need that for $z = 0$,

$$\frac{\partial U}{\partial x_r} = \frac{\partial U}{\partial y_r} = \begin{cases} -\sqrt{N}\nu & \text{for } r = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The Hessian is found to be of the form

$$D^2\tilde{U}(Q_0) = \begin{pmatrix} \mathbf{B} + \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} - \mathbf{C} \end{pmatrix}$$

with

$$\mathbf{B} + \mathbf{C} = \begin{pmatrix} c_1 + \kappa & 0 & \cdots & 0 & -\frac{\kappa}{\sqrt{N}}(N+\kappa) \\ 0 & c_1 + \kappa & \cdots & c_3 + \kappa & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & c_3 + \kappa & \cdots & c_1 + \kappa & 0 \\ -\frac{\kappa}{\sqrt{N}}(N+\kappa) & 0 & \cdots & 0 & \frac{\kappa}{N}(c_1+\kappa)(N+\kappa) \end{pmatrix}.$$

Since $\mathbf{B} - \mathbf{C}$ has the same form with the cross diagonal terms having the opposite sign, it suffices to investigate where the above matrix is singular and with it the Hessian of $\tilde{U}$. The determinant decomposes nicely into 2×2 subdeterminants plus into a single term in case the two diagonals cross when N is even. These subdeterminants are zero exactly at those values of κ given in (5.5) where the stability changes. The following theorem now follows easily.

THEOREM 6.1. *The Hessian of the reduced potential function $\tilde{U}$ is degenerate when the spectral stability of the ring with a central vortex changes, that is, for $\kappa = \kappa(r, N)$ with $2 \le r \le (N+2)/2$. The ordering of these critical values is*

$$-0.5 = \kappa(3, N) < \kappa(4, N) < \cdots < \kappa(2, N) = (N-1)^2/4.$$

Whenever κ is varied and passes through one of the critical values $\kappa(r, N)$ two pairs of exponents come from the imaginary axis, meet at the origin, and then one pair moves to the positive real axis, the other pair to the negative axis. The only exception is the case $r = (N+2)/2$ when N is even, where we have just one pair moving from the imaginary to the real axis.

Despite the similarities between the Newtonian potential and the logarithmic potential in the vortex problem, the results concerning the stability and bifurcation from a regular polygon configuration are different. In the Newtonian case the change in stability and the bifurcation occur for different values of the central mass. For the rotating vortices, these phenomena happen at the same value for the central vortex. It makes this problem more degenerate, but it also allowed the complete analysis of the spectral stability of regular polygon configurations, except in the case of the heptagon, which will be considered next.

7. The stability of the Thomson heptagon. For the heptagon configuration of Thomson, that is, for $N = 7$ with no central vortex, we now carry out the expansion of the Hamiltonian function as it was outlined in section 3. With the notation of that section we have

$$k(1,1) = 1, \qquad k(3,6) = k(4,5) = k(4,5) = k(6,3) = 2,$$

and for the quadratic part of U_1 we have

$$U_1^{(2)} = \frac{1}{2}\mathrm{Re}\left(\sum_{r+s=2 \bmod 7} \frac{7k(r,s) - rs}{2} z_r z_s \right),$$

so that

$$U_1^{(2)} = \frac{1}{2}\mathrm{Re}\left(3z_1^2 - 4z_3z_6 - 6z_4z_5\right).$$

In real variables, $z_j = x_j + \sqrt{-1}\; y_j$, $j = 1, \dots, 6$, we have then

$$U_1^{(2)} = \frac{1}{2}\left(3(x_1^2 - y_1^2) - 4(x_3x_6 - y_3y_6) - 6(x_4x_5 - y_4y_5)\right).$$

For the cubic terms, we notice that the values of $k(r,s,t)$ are given by $k(1,1,1) = 1$, $k(r,s,t) = 2$ if $r + s + t = 10$ and $k(r,s,t) = 26$ if $r + s + t = 17$. Then we find that for the homogeneous cubic part of U_1,

$$U_1^{(3)} = \frac{-1}{\sqrt{7}}\mathrm{Re}\left(z_1^3 - 4z_1z_3z_6 - 6z_1z_4z_5 - 5z_2^2z_6 - 16z_2z_3z_5 - 9z_2z_4^2 - 11z_3^2z_4 + z_5z_6^2\right).$$

For the fourth-degree terms, we first compute $k(r,s,t,u)$ (recall that $rstu \neq 0$): $k(1,1,1,1) = 1$, and when $r + s + t + u = 11$,

$$\begin{aligned} k(1,1,3,6) &= k(1,1,4,5) = k(1,2,2,6) = k(1,2,3,5) = k(1,2,4,4) \\ &= k(1,3,3,4) = k(2,2,2,5) = k(2,2,3,4) = k(2,3,3,3) = 2, \end{aligned}$$

while for $r + s + t + u = 18$ we have the possibilities

$$\begin{array}{lll} k(1,5,6,6) = 26, & k(2,4,6,6) = 42, & k(2,5,5,6) = 44, \\ k(3,3,6,6) = 48, & k(3,4,5,6) = 54, & k(3,5,5,5) = 57, \\ k(4,4,4,6) = 58, & k(4,4,5,5) = 62, & \end{array}$$

and these values are the same for $k(r,s,t,u)$ under a permutation of (r,s,t,u).

Then, for the coefficients of $C_{rstu} = C_{x_rx_sx_tx_u} = C_{y_ry_sy_ty_u} = -C_{x_rx_sy_ty_u}$ of the Taylor expansion of U_1, we have

$$C_{rstu} = \frac{3}{7}(7\;k(r,s,t,u) - rstu).$$

We find that

$$\begin{aligned}U_1^{(4)} = \frac{3}{7}\mathrm{Re}\biggl(&\frac{1}{4}z_1^4 - 2z_1^2z_3z_6 - 3z_1^2z_4z_5 - 5z_1z_2^2z_6 - 16z_1z_2z_3z_5 - 9z_1z_2z_4^2\\ &-11z_1z_3^2z_4 + z_1z_5z_6^2 - \frac{13}{3}z_2^3z_5 - 17z_2^2z_3z_4 - \frac{20}{3}z_2z_3^3 + 3z_2z_4z_6^2 + 4z_2z_5^2z_6\\ &+3z_3^2z_6^2 + 18z_3z_4z_5z_6 + 4z_3z_5^3 + \frac{11}{3}z_4^3z_6 + \frac{17}{2}z_4^2z_5^2\biggr).\end{aligned}$$

The complex variables $z_j = x_j + \sqrt{-1}y_j$ were useful for finding the expansion, but for the system of Hamiltonian differential equations we have to use the position variables x_j and the momenta y_j for $j = 0, 1, \ldots, 6$. After adding the terms, due to the rotation with $\nu = -3$, the Hamiltonian function (2.1) reads

$$V = \frac{3}{2}\sum_{j=0}^{6}(x_j^2 + y_j^2) + U_1^{(2)} + U_1^{(3)} + U_1^{(4)} + \cdots.$$

Since x_0 and y_0 occur only in the sum and setting $x_0 = y_0 = 0$ corresponds to fixing the center of vorticity at the origin, we can ignore these two variables from now on. The quadratic terms for the heptagon are therefore

$$\begin{aligned}V_2 = {}& 3x_1^2 + \frac{3}{2}(x_2^2 + y_2^2)\\ &+\frac{3}{2}x_3^2 - 2x_3x_6 + \frac{3}{2}x_6^2 + \frac{3}{2}x_4^2 - 3x_4x_5 + \frac{3}{2}x_5^2\\ &+\frac{3}{2}y_3^2 + 2y_3y_6 + \frac{3}{2}y_6^2 + \frac{3}{2}y_4^2 + 3y_4y_5 + \frac{3}{2}y_5^2.\end{aligned}$$

The quadratic terms can be simplified further by the rotation used already with (5.2) to give

$$V_2 = 3x_1^2 + \frac{3}{2}(x_2^2 + y_2^2) + \frac{5}{2}x_3^2 + 3x_4^2 + \frac{1}{2}x_6^2 + \frac{1}{2}y_3^2 + 3y_5^2 + \frac{5}{2}y_6^2.$$

In order to normalize the higher-order terms of the Hamiltonian function the linearized system should be in diagonal form as much as possible and we need one more symplectic transformation to complex-valued position and momenta variables given by

$$\begin{aligned}
x_1 &= \xi_1, & y_1 &= i\eta_1,\\
x_2 &= \tfrac{1}{\sqrt{2}}(\xi_2 + \eta_2), & y_2 &= \tfrac{i}{\sqrt{2}}(-\xi_2 + \eta_2),\\
x_3 &= \tfrac{1}{\sqrt{2\sqrt{5}}}(\xi_3 + \eta_3), & y_3 &= i\sqrt{\tfrac{\sqrt{5}}{2}}(-\xi_3 + \eta_3),\\
x_4 &= \xi_4, & y_4 &= i\eta_4,\\
x_5 &= \xi_5, & y_5 &= i\eta_5,\\
x_6 &= \sqrt{\tfrac{\sqrt{5}}{2}}(\xi_6 + \eta_6), & y_6 &= \tfrac{i}{\sqrt{2\sqrt{5}}}(-\xi_6 + \eta_6).
\end{aligned}$$

All momenta were changed so that the transformation is symplectic with multiplier $\sqrt{-1}$. The resulting Hamiltonian function is then written in a form so that the Lie transformation of Deprit can be applied:

$$H = -i\left(H_0 + H_1 + \frac{1}{2!}H_2 + \frac{1}{3!}H_3 + \cdots\right), \tag{7.1}$$

where H_k consists of homogeneous polynomials of degree $k+2$. The quadratic terms are given by

$$H_0 = 3\xi_1^2 + 3\xi_2\eta_2 + \sqrt{5}\xi_3\eta_3 + 3\xi_4^2 - 3\eta_5^2 + \sqrt{5}\xi_6\eta_6.$$

The higher-order terms have to be subjected to the same sequence of transformations. It results in 110 nonzero terms of degree three in H_1 and 595 of these of degree four in H_2. The Lie transformation lets us eliminate all terms that are in the range of the operator

$$\begin{aligned}
L_W H_0 &= \sum_{j=1}^{6} \frac{\partial H_0}{\partial \xi_j}\frac{\partial W}{\partial \eta_j} - \frac{\partial H_0}{\partial \eta_j}\frac{\partial W}{\partial \xi_j} \\
&= 6\left(\xi_1\frac{\partial W}{\partial \eta_1} + \xi_4\frac{\partial W}{\partial \eta_4} + \eta_5\frac{\partial W}{\partial \xi_5}\right) + 3\left(\eta_2\frac{\partial W}{\partial \eta_2} - \xi_2\frac{\partial W}{\partial \xi_2}\right) \\
&\quad + \sqrt{5}\left(\eta_3\frac{\partial W}{\partial \eta_3} - \xi_3\frac{\partial W}{\partial \xi_3} + \eta_6\frac{\partial W}{\partial \eta_6} - \xi_6\frac{\partial W}{\partial \xi_6}\right).
\end{aligned}$$

Due to the degeneracy of the problems with respect to the variables with indices one, four, and five, and due to the resonance between the variables with indices three and six, the complement of the range of the above operator is fairly large. The second author used his program called POLYPACK for the efficient manipulation of polynomials in several variables to carry out this normalization by machine. The calculations were performed with the coefficients stored as floating point numbers. After having obtained the answer, the computations were then checked in part with the help of *Mathematica*.

Without going into the details of these computations, the results were that at order three the following four terms were left over in the complex form of H_1:

$$H_1 = 2\sqrt{10/7}\left[(-\xi_3\eta_3 + \xi_6\eta_6)\xi_5 + (\xi_6\eta_3 - \xi_3\eta_6)\eta_4\right].$$

At the next order there were 46 nonzero terms left over, and there are too many to be listed here. Since the main purpose of the normalization is to investigate the stability of the origin, we will use polar variables where appropriate and return to the real variables in the other cases:

$$\begin{aligned}
&\xi_j = r_j e^{i\phi_j}, \quad && \eta_j = r_j e^{-i\phi_j} \quad && \text{for } j = 2, 3, \text{ and } 6, \\
&\xi_j = x_j, && \eta_j = iy_j && \text{for } j = 1, 4, \text{ and } 5.
\end{aligned}$$

The quadratic part of the Hamiltonian for the heptagon is then

$$H_0 = 3x_1^2 + 3r_2^2 + \sqrt{5}(r_3^2 + r_6^2) + 3x_4^2 + 3y_5^2.$$

The function is obviously indefinite as far as the variables y_1, y_4, and x_5 are concerned. The third-order terms in the normal form are then

$$H_1 = \sqrt{10/7}\left[2x_5(r_6^2 - r_3^2) - y_4 r_3 r_6 \sin(\phi_6 - \phi_3)\right].$$

They do not exclude Liapunov stability of the origin since they are dominated by the quadratic terms $\sqrt{5}(r_3^2 + r_6^2)$ near the origin. Therefore, we have to look at terms of degree four in the normalized H_2. All terms containing a factor r_2^2, r_3^2, or r_6^2 are of no interest since they cannot change the positive definiteness of the Hamiltonian near the origin. The terms in the normalized $H_2/2!$, which do not meet that criteria, are

$$\frac{9}{4}(x_5^2 + y_4^2)^2 - 17(x_4 x_5 + y_4 y_5)^2.$$

Another term with $x_5 y_4(x_4 x_5 + y_4 y_5)$ could have appeared in the normal form, but due to the symmetry of our problem its coefficient turns out to be zero. Since the coefficient of the first term, that is, of $(x_5^2 + y_4^2)^2$, is positive, we cannot make the Hamiltonian negative by choosing any values of x_5 and y_4 near zero. The coefficient of the second term is negative, but it cannot make the entire Hamiltonian negative near the origin, since each of its terms is dominated by the quadratic terms $3x_4^2 + 3y_5^2$ in H_0. No information is available with respect to the variable y_1, and we have to expect that the Hamiltonian remains indefinite in that direction. Actually this could have been predicted since fixing y_1 selects one of the relative equilibria from the rotational symmetry. Except for the extensive numerical computations by machine we have thus shown the following theorem.

THEOREM 7.1. *Except for the rotational symmetry the Thomson heptagon is locally Liapunov stable.*

The result corresponds to the one of Mertz in [6], which was obtained by a completely different method. It also agrees with what is observed when a careful numerical integration of the problem is performed. It differs from the results in [2], which predicted instability with the appearance of third-order terms in the normalized Hamiltonian function. To be precise the normalization for the heptagon is not given in that paper, and the statement about instability concerns a Hamiltonian whose quadratic part can be diagonalized. As shown above the appearance of terms of third degree in the Hamiltonian does not necessarily destroy the local stability of the origin. The effect of the third-order terms on the global stability is, of course, another issue.

REFERENCES

[1] H. AREF, *On the equilibrium and stability of a row of point vortices*, J. Fluid Mech., 290 (1995), pp. 167–181.

[2] L. G. KHAZIN, *Regular polygons of point vortices and resonance instability of steady states*, Soviet Phys. Dokl., 21 (1976), pp. 567–569.

[3] G. KIRCHHOFF, *Mechanik*, Vorlesungen über Mathematische Physik, Teubner, Leipzig, 1883.

[4] D. LEWIS AND T. RATIU, *Polygonal vortex configurations*, in New Trends for Hamiltonian Systems and Celestial Mechanics, Vol. 8, E. A. Lacomba and J. Llibre, eds., World Scientific, Singapore, 1996, pp. 249–262.

[5] J. C. MAXWELL, *Stability of the motion of Saturn's rings*, in Maxwell on Saturn's Rings, S. Brush, C. W. F. Everitt, and E. Garber, eds., MIT Press, Cambridge, MA, 1983.

[6] G. J. MERTZ, *Stability of body-centered polygonal configurations of ideal vortices*, Phys. Fluids, 21 (1978), pp. 1092–1095.

[7] K. R. MEYER AND D. S. SCHMIDT, *Bifurcations of relative equilibria in the N-body and Kirchhoff problems*, SIAM J. Math. Anal., 19 (1988), pp. 1295–1313.

[8] R. MOECKEL, *Linear stability analysis of some symmetrical classes of relative equilibria*, in Hamiltonian Dynamical Systems, History, Theory, and Applications, H. S. Dumas, K. R. Meyer, and D. S. Schmidt, eds., IMA Vol. Math. Appl. 68, Springer, New York, 1994, pp. 291–317.

[9] G. K. MORIKAWA AND E. V. SWENSON, *Interacting motion of rectilinear geostrophic vortices*, Phys. Fluids, 14 (1971), pp. 1058–1073.

[10] W. OLVA, *The motion of two dimensional vortices with mass as a singular perturbation Hamiltonian problem*, in New Trends for Hamiltonian Systems and Celestial Mechanics, Vol. 8, E. A. Lacomba and J. Llibre, eds., World Scientific, Singapore, 1996, pp. 301–310.

[11] J. I. Palmore, *Measure of degenerate relative equilibria* I, Ann. Math., 140 (1976), pp. 421–429.

[12] D. S. Schmidt, *Spectral stability of relative equilibria in the $N + 1$ body problem*, in Hamiltonian Systems and Celestial Mechanics, E. A. Lacomba and J. Llibre, eds., World Scientific, Singpore, 1995, pp. 321–341.

[13] J. J. Thomson, *A Treatise on the Motion of Vortex Rings: An Essay to which the Adams Prize was Adjusted in* 1882, University of Cambridge, Macmillan, London, 1883.

SIAM J. Math. Anal.
Vol. 31, No. 2, pp. 251–273

A COMPREHENSIVE MATHEMATICAL MODEL FOR A MULTISPECIES FLOW THROUGH GROUND COFFEE[*]

A. FASANO[†] AND F. TALAMUCCI[†]

Abstract. A model of filtration in a multispecies porous medium with mechanical and chemical interaction between the flow and the porous matrix is presented. The species removed by the medium are transported either as solids or as solutes. The accumulation of the migrating particles in proximity of the outflow surface gives rise to the growth of a compact layer, with high hydraulic resistance.

The corresponding mathematical problem is a free boundary problem for parabolic and hyperbolic partial differential equations. Existence and uniqueness globally in time are proved.

Key words. nonstandard filtration processes in porous media, systems of PDEs in a free boundary domain, convection and diffusion

AMS subject classifications. 35R35, 76S05, 76T05, 76R50.

PII. S0036141098336698

1. The physical problem.

This paper is in the framework of a research which was promoted years ago by illycaffè s.p.a. (Trieste, Italy) aimed at describing the extremely complex set of phenomena taking place during the filtration of water through ground coffee in the espresso-coffee machine. For a review of the main results previously obtained on the espresso-coffee problem and other related models, see [6]. The studies performed so far have pointed out various aspects of the process, whose analysis has been carried out separately due to the relevant difficulties of the corresponding mathematical models. The present paper is the conclusive effort in providing a comprehensive description of the process, taking into account mechanical and chemical interactions between the flow and the porous matrix.

Our aim is to combine the model presented in [9] for the removal and transport of a single family of particles building up a low conductivity compact layer, and the models of [8], [13], in which several species are transported by the flow (either as solutes or as solid particles) but are allowed to leave the system. At the same time we try to keep the model as close as possible to reality, removing some simplifying assumptions introduced in previous investigations (see, e.g., the choice of a uniform and constant porosity in [9]).

Here we will not deal with the first stage of the process, in which the initially dry medium is penetrated by water. We recall however that invasion problems have been considered rather extensively in this context [7], [10], [4]. In addition we remark that in this paper we allow for nonuniform distribution of the removable species at the time $t = 0$ (a nontrivial extension of [9]), so that the initial conditions could be those determined at the end of the invasion process. One feature of [8] not incorporated here is the deformability of the porous matrix, since there is very little hope to obtain any experimental information on it. However, this would not represent a

[*] Received by the editors April 2, 1998; accepted for publication (in revised form) February 10, 1999; published electronically December 21, 1999. This work was supported by the Istituto Nazionale di Alta Matematica Francesco Severi and by the Research Contract between E.C.M.I and illycaffè s.p.a.

http://www.siam.org/journals/sima/31-2/33669.html

[†] Dipartimento di Matematica "U. Dini," Viale Morgagni, 67/a 50134 Firenze, Italy (fasano@math.unifi.it, talamucci@math.unifi.it).

major complication from a mathematical point of view. More could be done in the field on nonisothermal flows, in the spirit of the studies performed in [1], [2] for flow processes in the manufacturing of composite materials.

Let us come to the description of our paper.

We are going to formulate a mathematical model for a process of multispecies flow through a saturated porous medium occupying a layer of given thickness. A positive pressure is imposed at the inflow surface and during the filtration through the pores some components of the layer are removed by the action of the flow. The main features of the model are the following:

(i) The flow is able to extract mass in various ways from the porous matrix. The removed components are both sufficiently fine solid particles, which are transported convectively by the flow (and eventually accumulated in a compactlayer), and substances which may also diffuse in the liquid (typically as solutes).

(ii) The porosity is affected by the removal process, but here we neglect additional effects such as mechanical compression of the porous matrix by the flow.

(iii) The removal rate depends on the concentration of the particles still bound to the porous matrix and on the liquid flux intensity.

(iv) The hydraulic permeability in the part of the layer where the removal process takes place depends on the concentration of each species and on the porosity, while it is much lower and constant in the compact layer.

(v) The fine particles do not leave the system but they accumulate in the vicinity of the outflow surface, giving rise to a compact layer, whose structure depends on the history of the flow.

(vi) Effects of gravity and interdiffusion can be neglected.

In sections 2 and 3 we will introduce the mathematical model corresponding to (i)–(vi). Under the hypotheses listed in section 4 we will show existence and uniqueness for the solution of the mathematical problem, provided that the dependence of the removal rate on the particle's concentrations and on the liquid flux (cf. point (iii)) is sufficiently mild. Finally, in section 6 we will discuss some qualitative properties of the solution.

2. The equations governing the displacement of each species.

The model we are going to present may look exceedingly complicated because of the large number of quantities involved and the complexity of the phenomena taking place simultaneously. Nevertheless, as we shall see, it is not too difficult to find a way through it, although the existence proof is necessarily technical, since the basic unknowns are just the overall volumetric velocity (as a function of time and space) and the thickness of the compact layer. All the other quantities can be calculated using these two unknowns as input.

Different from previous attempts, based on heuristic simplifications, and therefore of limited applicability, here we set up a completely rigorous model. In such a way, not only do we have a reliable description of the whole process, but we also provide a basis for the analysis of all physical situations involving flows through porous media with movable or soluble components.

First, we define the concentrations of the various components entering the process.

As in [13], we denote by b_i and m_i the concentration of the i-species, $i = 1, \dots, n$, (mass per unit volume of the total system) when it is bound to the porous matrix and when it is moving, respectively.

After defining $\rho_i^{(\ell)}$, $\ell = b, m$, as the densities of each species (i.e., mass per unit volume of that component), we find it convenient to introduce also the specific volumes of the species in the system:

$$\begin{cases} \theta_i = b_i/\rho_i^{(b)}, & i = 1, \dots, n, \\ \eta_i = m_i/\rho_i^{(m)}, & i = 1, \dots, n, \\ \eta_w = m_w/\rho_w & \text{water.} \end{cases} \tag{2.1}$$

We introduce the following distinction among the components:

- $b_i, i = 1, \dots, k < n$, refers to fine solid particles bound to the porous matrix (typical diameters $1 - 10\,\mu$);
- $m_i, i = 1, \dots, k < n$, are the concentrations of the particles moving in the flow;
- $b_i, i = k + 1, \dots, n$, refers to substances in the porous matrix which once removed are dissolved in the fluid as solutes or as droplets;
- $m_i, i = k + 1, \dots, n$, refers to dissolved substances.

Let us denote by V_i and V_w the molecular velocities of the species m_i and of the water, respectively.

We define the following quantities:

$$\varepsilon = \eta_{(s)} + \eta_w \quad \text{porosity}, \qquad \eta_{(s)} = \sum_{i=k+1}^{n} \eta_i, \tag{2.2}$$

$$q = \eta_w V_w + \sum_{i=k+1}^{n} \eta_i V_i \quad \text{volumetric compound velocity.} \tag{2.3}$$

The meaning of (2.2) is that since the medium is constantly saturated, the pore volume is entirely occupied by the fluid and the partition between the volume fractions of water (η_w) and solutes ($\eta_{(s)}$) is emphasized. The necessity of introducing (2.3) comes from the fact that q is precisely the quantity which obeys Darcy's law (see (3.3)). We also define the cumulative volume fractions

$$\eta_{(p)} = \sum_{i=1}^{k} \eta_i, \quad \theta_{(p)} = \sum_{i=1}^{k} \theta_i, \quad \theta_{(s)} = \sum_{i=k+1}^{n} \theta_i, \tag{2.4}$$

and $\theta_0 = \theta_0(x)$ as the volume fraction (with respect to the unit volume of the total system) of the rigid porous skeleton. The part of unit volume complementary to (2.2) and occupied by solid components is partitioned as follows:

$$1 - \varepsilon = \eta_{(p)} + \theta_0 + \theta_{(s)} + \theta_{(p)}, \tag{2.5}$$

where we see the contribution of the nondeformable matrix (θ_0) of the solid moving ($\eta_{(p)}$) or movable ($\theta_{(p)}$) particles and of the soluble substances at the solid state ($\theta_{(s)}$).

As in [13], we start from the conservation laws of each species:

$$\frac{\partial m_i}{\partial t} + \frac{\partial}{\partial x}(m_i V_i) = -\frac{\partial b_i}{\partial t}, \ i = 1, \dots, n, \tag{2.6}$$

$$\frac{\partial m_w}{\partial t} + \frac{\partial}{\partial x}(m_w V_w) = 0. \tag{2.7}$$

The boundary $s(t)$, whose position is unknown a priori, separates the region $0 < x < s(t)$, where the removal process occurs, from the region $s(t) < x < 1$, where the particles accumulate, forming the compact layer.

Arguing in the same way as in [12], we see that the mass balances at the boundary $s(t)$, obtained by integration of (2.6) and (2.7) are

$$[[m_i + b_i]]\,\dot{s}(t) - [[m_i V_i]] = 0, \quad i = 1, \dots, n, \tag{2.8}$$

$$[[m_w]]\,\dot{s}(t) - [[m_w V_w]] = 0, \tag{2.9}$$

where $[[\chi]]$ (here and throughout the paper) denotes the jump of χ at $s(t)$:

$$[[\chi]] = \lim_{x \to s(t)^+} \chi(x,t) - \lim_{x \to s(t)^-} \chi(x,t). \tag{2.10}$$

Denoting by

$$V = q/\varepsilon$$

the average velocity of the liquid and assuming (cf. [6], [13]) that

$$V_i = \alpha_i V, \quad 0 < \alpha_i \le 1,\ i = 1, \dots, k, \tag{2.11}$$

where α_i are slowing factors due to the shocks, we get from (2.8)

$$[[m_i + b_i]]\,\dot{s}(t) - \alpha_i \left[\left[\frac{q}{\varepsilon} m_i\right]\right] = 0,\ i = 1, \dots, k. \tag{2.12}$$

On the other hand, for the dissolved species $i = k+1, \dots, n$ and for water we introduce the molecular diffusive flux with respect to the average (i.e., over all species) velocity $V = q/\varepsilon$:

$$J_i = \frac{m_i}{\varepsilon}(V_i - V),\ i = k+1, \dots, n, \quad J_w = \frac{m_w}{\varepsilon}(V_w - V). \tag{2.13}$$

The diffusive flux (2.13) is related to the concentration of the species in the liquid in the following way (cf., e.g., [5]):

$$J_i = -D_i \nabla(m_i/\varepsilon), \quad i = k+1, \dots, n, \tag{2.14}$$

where D_i are the diffusion coefficients and m_i/ε represents concentrations in the flow. In (2.14) interdiffusion is neglected, according to assumption (vi) of section 1.

From (2.13) and (2.14) we obtain the following expression for the partial mass fluxes $m_i V_i$:

$$m_i V_i = m_i V - \varepsilon D_i \frac{\partial}{\partial x}\left(\frac{m_i}{\varepsilon}\right),\ i = k+1, \dots, n. \tag{2.15}$$

The concentration of the i-species, $i = k+1, \dots, n$, in the flowing liquid and the concentration b_i, $i = k+1, \dots, n$, are supposed to be continuous across the boundary s:

$$\left[\left[\frac{m_i(s(t),t)}{\varepsilon(s(t),t)}\right]\right] = [[b_i(s(t),t)]] = 0,\ i = k+1, \dots, n. \tag{2.16}$$

It is worth noticing that by summing up (2.8), $i = k+1, \dots, n$, and (2.9) and recalling (2.1)–(2.3) and (2.16) we get the following condition at $s = s(t)$:

$$[[\varepsilon]]\,\dot{s}(t) = [[q]]\,. \tag{2.17}$$

(If ε jumps, q has to jump in order to preserve saturation.) Concerning the migrating particles, if we assume, as it appears quite natural, that the solid particles are no longer mobile in the compact layer, i.e.,

$$m_i(s(t)^+, t) = 0, \quad i = 1, \dots, k, \tag{2.18}$$

then mass balances (2.12) become

$$\begin{aligned} &\{b_i(s(t)^+, t) - \big(m_i(s(t)^-, t) + (b_i(s(t)^-, t))\big)\}\dot{s}(t) \\ &= -\alpha_i \frac{q(s(t), t)}{\varepsilon(s(t), t)} m_i(s(t)^-, t), \ i = 1, \dots, k. \end{aligned} \tag{2.19}$$

Let us denote now by M_i, $i = 1, \dots, k$, the concentration of the i-species in the compact layer $s(t) < x \leq 1$.

If (2.18) holds throughout the compact layer (i.e., $M_i = b_i$ in $s(t) < x < 1$, $i = 1, \dots, k$), we obviously have

$$\frac{\partial M_i}{\partial t}(x, t) = 0, \quad i = 1, \dots, k, \ s(t) < x < 1, \tag{2.20}$$

but the functions $M_i(x)$ are not known. The structure of the compact layer, corresponding to a sequence $M_1, \dots, M_k$ depends on the history of the process, since it depends on the incoming flux of particles.

In order to model the formation of the compact layer, a constraint f for the concentrations M_i must be assigned, defining a *packing configuration*

$$f(M_1, \dots, M_k) = 0, \tag{2.21}$$

f being a C^1 function such that $\partial f / \partial M_i > 0$, $i = 1, \dots, k$.

Since in (2.19) we identify M_i with $b_i(s(t)^+, t)$, $i = 1, \dots, k$, from (2.21), we get

$$f\left(-\frac{1}{\dot{s}}\alpha_1 \frac{q}{\varepsilon} m_1^- + m_1^- + b_1^-, \cdots, -\frac{1}{\dot{s}}\alpha_k \frac{q}{\varepsilon} m_k^- + m_k^- + b_k^-\right) = 0, \tag{2.22}$$

where r^- stands for $r(s(t)^-, t)$. The solvability of (2.22) with respect to $\dot{s}$ is guaranteed by the condition

$$\frac{q}{\varepsilon} \sum_{i=1}^{k} \alpha_i m_i^- \frac{\partial f}{\partial M_i} \neq 0. \tag{2.23}$$

A simple but reasonable way to prescribe the packing configuration (2.21) is the following, which refers to the specific volumes of the species:

$$\sum_{i=1}^{k} \frac{M_i}{\rho_i} = \Theta, \tag{2.24}$$

where ρ_i is the density of M_i and Θ is a known quantity. Equation (2.24) shows that the layer is compact when the incoming particles occupy the maximum specific volume at their disposal, represented by Θ.

If only one species of fine particles is present (i.e., $k = 1$), condition (2.24) implies the knowledge of the particle concentration in the compact layer, as in [9].

3. The complete mathematical model. The removal process (species $i = 1, \ldots, k$) occurs only in the region $D_T = \{(x,t) : 0 < x < s(t),\ 0 < t < T\}$. On the other hand, the extraction process (species $i = k+1, \ldots, n$) may take place also in the compact layer $R_T = \{(x,t) : s(t) < x < 1,\ 0 < t < T\}$. The governing equations, which will be noted just below (together with the new symbols introduced), are

$$\frac{\partial m_i}{\partial t} + \frac{\partial}{\partial x}\left(\alpha_i m_i \frac{q}{\varepsilon}\right) = -\frac{\partial b_i}{\partial t},\ i = 1, \ldots, k,\ (x,t) \in D_T, \tag{3.1}$$

$$\begin{aligned}&\frac{\partial m_i}{\partial t} + \frac{\partial}{\partial x}\left(-D_i \varepsilon \frac{\partial}{\partial x}\frac{m_i}{\varepsilon}\right) + \frac{\partial}{\partial x}\left(m_i \frac{q}{\varepsilon}\right) = -\frac{\partial b_i}{\partial t},\\ &i = k+1, \ldots, n \quad (x,t) \in D_T \cup R_T,\end{aligned} \tag{3.2}$$

$$q = -K(b, m, \varepsilon)\frac{\partial p}{\partial x}, \quad (x,t) \in D_T \cup R_T, \tag{3.3}$$

$$\frac{\partial b_i}{\partial t} = -F_i(q,b)G_i[b_i - \beta_i(q,b)]^+, \quad i = 1, \ldots, k, \quad (x,t) \in D_T, \tag{3.4}$$

$$\frac{\partial b_i}{\partial t} = -H_i(q,b), \quad i = k+1, \ldots, n, \quad (x,t) \in D_T \cup R_T, \tag{3.5}$$

$$\frac{\partial \varepsilon}{\partial t} + \frac{\partial q}{\partial x} = -\frac{\partial}{\partial t}\theta_{(s)}, \quad (x,t) \in D_T \cup R_T, \tag{3.6}$$

$$\frac{\partial q}{\partial x} + \frac{\partial}{\partial x}\sum_{i=1}^{k}\left(\alpha_i \eta_i \frac{q}{\varepsilon}\right) = 0, \quad (x,t) \in D_T \cup R_T, \tag{3.7}$$

together with the initial and boundary conditions

$$m_i(x,0) = m_{i,0}(x), \quad i = 1, \ldots, n, \quad x \in [0,1], \tag{3.8}$$

$$\varepsilon(x,0) = \varepsilon_0(x), \quad x \in [0,1], \tag{3.9}$$

$$b_i(x,0) = b_{i,0}(x), \quad i = 1, \ldots, n, \quad x \in [0,1], \tag{3.10}$$

$$m_i(0,t) = 0, \quad i = 1, \ldots, k, \quad 0 \le t \le T, \tag{3.11}$$

$$\begin{aligned}&D_i \varepsilon(0,t)\frac{\partial m_i}{\partial x}(0,t) = \frac{q(0,t)m_i(0,t)}{\varepsilon(0,t)},\\ &i = k+1, \ldots, n,\ 0 < t < T,\end{aligned} \tag{3.12}$$

$$[[p]] = 0, \quad x = s(t), \tag{3.13}$$

$$\left[\left[\frac{m_i}{\varepsilon}\right]\right] = 0, \quad x = s(t), \quad i = k+1, \ldots, n, \tag{3.14}$$

$$\left[\left[D_i \varepsilon \frac{\partial}{\partial x}\frac{m_i}{\varepsilon}\right]\right] = 0, \quad x = s(t), \quad i = k+1, \ldots, n, \tag{3.15}$$

$$\big(\Theta - (\eta_{(p)} + \theta_{(p)})\big)\,\dot{s} = -\frac{q}{\varepsilon}\sum_{i=1}^{k}\alpha_i \eta_i, \quad x = s(t), 0 \le t \le T, \tag{3.16}$$

$$\frac{\partial}{\partial x}\frac{m_i}{\varepsilon}(1,t) = 0, \quad i = k+1, \ldots, n, \tag{3.17}$$

$$p(0,t) = p_0(t) > 0, \quad p(1,t) = 0, \quad 0 \le t \le T, \tag{3.18}$$

$$s(0) = 1. \tag{3.19}$$

The unknown quantities for (3.1)–(3.19) are the concentrations $b_i(x,t)$, $m_i(x,t)$, $i = 1, \ldots, n$ (or the specific volumes θ_i, η_i, $i = 1, \ldots, n$; see (2.1)), the liquid flux $q(x,t)$, the porosity $\varepsilon(x,t)$, the liquid pressure $p(x,t)$, and the free boundary $s(t)$.

Equations (3.1) and (3.2) are the mass balances (2.6) according to assumption (i) of section 1 with the specification of V_i (see (2.11) and (2.15)). Equation (3.3) gives the liquid flux q (see definition (2.3)), while (3.4) and (3.5) regulate the release of particles or substances from the porous matrix (see [6] and [8] for a more detailed explanation of the physical model). The functions $\beta_i(q, b)$ are threshold concentrations for the removal of the species $i = 1, \dots, k$ (fine particles). As pointed out in [9] they play a crucial role in explaining some qualitative features of the process.

In (3.3)–(3.5) by b and m we mean the vectors $(b_1, \dots, b_n)$ and $(m_1, \dots, m_n)$, respectively. We will also consider (3.4) and (3.5) in terms of volumetric contents by introducing the functions $\hat{\Phi}_i = \Phi_i/\rho_i(q, \rho_1\theta_1, \dots, \rho_n\theta_n)$ and $\hat{H}_i$ defined analogously and by considering the removal laws

$$\frac{\partial \theta_i}{\partial t} = -\hat{\Phi}_i(q, \theta), \quad i = 1, \dots, k, \tag{3.20}$$

$$\frac{\partial \theta_i}{\partial t} = -\hat{H}_i(q, \theta), \quad i = k+1, \dots, n. \tag{3.21}$$

The global conservation laws (3.6) and (3.7) have been introduced in [13], under the hypothesis (which will be assumed true from now on) $\rho_i^{(b)} = \rho_i^{(m)} = \rho_i, i = 1, \dots, n$. Equation (3.6), which expresses incompressibility and saturation of the system, is obtained by replacing in (2.7) (divided by ρ_w) the quantities η_w and $\eta_w V_w$ deduced from (2.2) and (2.3), respectively, and by taking into account (2.6) for $i = k+1, \dots, n$. There is an easy way of reading (3.6): the divergence of the flux must compensate the possible unbalance between the pore volume created by dissolution and the volume previously occupied in the solid matrix by the dissolved substances (the two extremes being $\varepsilon + \theta_{(s)} =$ constant, or $\partial\theta_{(s)}/\partial t$ negligible in comparison to $\partial\varepsilon/\partial t$, which is the approach used in [9]).

Equation (3.7) comes from summing up (2.6) (divided by ρ_i), for $i = 1, \dots, n$ and (2.7) (divided by ρ_w). The terms containing the time-derivatives eliminate by virtue of (2.2) and (2.5), and using (2.3) gives the final form (3.7). The balance (3.7) states that the global volumetric flux of all the moving components (water and species η_i, $i = 1, \dots, n$) is constant at each time t.

In writing (3.12) we assume that the velocity of the i-species, $i = k+1, \dots, n$ vanishes at $x = 0$ (cf. (2.15)). Equation (3.14) is (2.16), while (3.15) comes from (2.8), (2.15), (2.16), and (2.17) and expresses the continuity of the diffusive flux of each species at the interface. By condition (3.17) we assume that the effect of diffusion is negligible at the outflow surface $x = 1$. Finally, the free boundary condition (3.16) is obtained by assuming the configuration constraint (2.24) and by summing up (2.19) with respect to i.

Note that the function $\theta_0(x)$, which gives the volume fraction of the porous skeleton (cfr. (2.5)), can be calculated by means of (3.8)–(3.10):

$$\theta_0(x) = 1 - (\varepsilon_0(x) + \eta_{(k),0}(x) + \theta_{(k),0}(x) + \eta_{(s)_0}(x)). \tag{3.22}$$

Obviously, the initial given functions appearing on the right-hand side of (3.22) are physically consistent only if $0 < \theta_0(x) < 1$.

We conclude this section by remarking that the water volume fraction η_w and the water velocity V_w can be computed a posteriori by means of (2.2), (2.3), and (2.11) once problem (3.1)–(3.19) has been solved.

4. List of assumptions. We assume that $K(b, m, \varepsilon)$, $F_i(q, b)$, $G_i(\eta)$, $\beta_i(q, b)$, $i = 1, \dots, k$, $H_i(q, b)$, $i = k+1, \dots, n$, are C^1-functions of their respective arguments and each first derivative is Lipschitz continuous. The given functions $\varepsilon_0(x)$, $m_{i,0}(x)$, $b_{i,0}(x)$, $i = 1, \dots, n$, and $p_0(t)$ are assumed to be C^1 with respect to their argument with bounded derivatives. Moreover, we will assume that there exist positive constants K_m, K_M, p_0^m, p_0^M, and m_0 such that

$$0 < K_m \le K(b, m, \varepsilon) \le K_M \quad \forall\, b, m, \varepsilon, \tag{4.1}$$

$$0 < p_0^m \le p_0(t) \le p_0^M, \quad t \ge 0. \tag{4.2}$$

The initial distribution of the various species must satisfy

$$m_{i,0}(x) \not\equiv 0, \quad b_{i,0}(x) \not\equiv 0, \quad x \in [0,1], \tag{4.3}$$

$$\eta_{(k),0}(x) + \theta_{(k),0}(x) < \Theta < 1 - (\theta_0(x) + {\theta_{(s)}}_0(x)), \quad x \in [0,1]. \tag{4.4}$$

Moreover, for simplicity we take the compatibility conditions

$$m_{i,0}(0) = 0, \quad i = 1, \dots, k. \tag{4.5}$$

Remark 4.1. Combining (4.4) with (2.5) and recalling (3.22) we get the following restriction for the initial data:

$$1 - \big(\Theta + \theta_0(x) + {\theta_{(s)}}_0(x)\big) < \varepsilon_0(x) < 1 - \theta_0(x), \quad x \in [0,1]. \tag{4.6}$$

We set

$$\begin{cases} \varepsilon_0^m = \min\limits_{x\in[0,1]} \big(1 - (\Theta + \theta_0(x) + {\theta_{(s)}}_0(x))\big)\,, \\ \varepsilon_0^M = \max\limits_{x\in[0,1]} (1 - \theta_0(x)) \end{cases} \tag{4.7}$$

(note that $0 < \varepsilon_0^m < \varepsilon_0^M < 1$). From (4.6) and (4.7), we deduce

$$0 < \varepsilon_0^m < \varepsilon_0(x) < \varepsilon_0^M < 1, \quad x \in [0,1]. \tag{4.8}$$

The functions introduced in the removal laws (3.4) and (3.5) have the properties

$$F_i(q, b_1, \dots, b_n) \ge 0,\ i = 1, \dots, k \quad \forall q \ge 0, 0 \le b_j \le b_{j,0},\ 1 \le j \le n, \tag{4.9}$$

$$\begin{aligned} &H_i(q, b_1, \dots, b_n) \ge 0,\ i = k+1, \dots, n \quad \forall q \ge 0, 0 \le b_j \le b_{j,0},\ 1 \le j \le n, \\ &H_i(q, 0, \dots, 0) = 0,\ i = k+1, \dots, n \quad \forall q \ge 0, \end{aligned} \tag{4.10}$$

$$\frac{\partial \beta_i}{\partial q}(q, b_1, \dots, n) \le 0,\ i = 1, \dots, k \quad \forall q \ge 0, 0 \le b_j \le b_{j,0},\ 1 \le j \le n. \tag{4.11}$$

Moreover, setting

$$\Phi = \sum_{i=1}^{k} \Phi_i = \sum_{i=1}^{k} F_i G_i, \quad H = \sum_{i=k+1}^{n} H_i \tag{4.12}$$

and defining the norm

$$\|\Psi\| = \sup_{q \ge 0, 0 \le b_i \le b_{i,0}} |\Psi(q, b)|, \tag{4.13}$$

we assume that

$$\|\Phi\| < \infty, \quad \|H\| < \infty. \tag{4.14}$$

For a function $g(\cdot)$, $\|g\|$ will denote the sup-norm. For a vector $v(t) = (v_1(t), \dots, v_n(t))$, we define

$$\|v\|_T = \max_{1\le i\le n} \sup_{0\le t\le T} |v_i(t)|. \tag{4.15}$$

We will denote by L_g^z the Lipschitz constant of the function g with respect to the variable z. (In the case $z = x$ we write simply L_g.) In particular, defining the vector

$$S = (S_1, \dots, S_n) = (\Phi_1, \dots, \Phi_k, H_{k+1}, \dots, H_n) \tag{4.16}$$

we will assume that positive constant values $L^q_{S_i}$, $L^b_{S_i}$, $L^q_{S_{i_q}}$, $L^b_{S_{i_{b_j}}}$, $i, j = 1, \dots, n$, exist such that

$$\begin{aligned}
&|S_i(q_1, b) - S_i(q_2, b)| \le L^q_{S_i}|q_1 - q_2|,\\
&|S_i(q, b^{(1)}) - S_i(q, b^{(2)})| \le L^b_{S_i} \max_{1\le j\le n} |b_j^{(1)} - b_j^{(2)}|,\\
&\left|\frac{\partial S_i}{\partial q}(q_1, b) - \frac{\partial S_i}{\partial q}(q_2, b)\right| \le L^q_{S_{i_q}}|q_1 - q_2|,\\
&\left|\frac{\partial S_i(q, b^{(1)})}{\partial b_j} - \frac{\partial S_i(q, b^{(2)})}{\partial b_j}\right| \le L^b_{S_{i_{b_j}}} \max_{1\le r\le n} |b_r^{(1)} - b_r^{(2)}|,
\end{aligned} \tag{4.17}$$

where q, q_1, q_2 are nonnegative real numbers and b, $b^{(1)}$, $b^{(2)}$ are vectors in $\mathbb{R}^n$ whose components belong to the intervals $[0, b_{j,0}]$, $j = 1, \dots, n$.

We introduce also the vector $\hat S$ defined as (cf. (3.20) and (3.21))

$$\hat S = (\hat S_1, \dots, \hat S_n) = (\hat\Phi_1, \dots, \hat\Phi_k, \hat H_{k+1}, \dots, \hat H_n). \tag{4.18}$$

We will denote by $L^q_{\hat S_i}$, $L^\theta_{\hat S_i}$, $L^q_{\hat S_i q}$, ${L_{\hat S_i \theta_j}}^\theta$, $i, j = 1, \dots, n$, the Lipschitz constant of $\hat S_i$, $\partial \hat S_i/\partial q$, $\partial \hat S_i/\partial \theta_j$ defined in the same way as in (4.17).

5. Existence and uniqueness of the solution. We will show the existence of solutions of (3.1)–(3.19) basing our proof on the Schauder's fixed point theorem. Our first aim is to calculate the volumetric flux for $t = 0$.

5.1. Determination of $q(x, 0)$. Let us integrate (3.3) with respect to x:

$$\int_0^{s(t)} \frac{q(\xi, t)}{K(b(\xi, t), m(\xi, t), \varepsilon(\xi, t))} d\xi = p_0(t) - p(s(t)^-, t). \tag{5.1}$$

We notice that the liquid flux in the compact layer is a function of time only as it can be deduced from (3.7). It will be denoted by $q_c(t)$. In the compact layer we have $q_c(t)(1 - s(t))/K_0 = p(s(t)^+, t)$ (see assumption (iv) of section 1, (2.20), and (3.18)), where K_0 is the hydraulic conductivity.

According to (3.13), we have

$$\int_0^{s(t)} \frac{q(\xi, t)}{K(b(\xi, t), m(\xi, t), \varepsilon(\xi, t))} d\xi = p_0(t) - (1 - s(t))\,\frac{q_c(t)}{K_0}. \tag{5.2}$$

Define now

$$l(x, t) = \frac{\displaystyle\sum_{i=1}^{k} \alpha_i \eta_i(x, t)}{\varepsilon(x, t)}, \quad l_0(x) = l(x, 0). \tag{5.3}$$

Equation (3.7) yields

$$q(x,t) = f(t)\frac{1}{1+l(x,t)}, \quad 0 \le x \le s(t), \tag{5.4}$$

where $f(t) = q(0,t)$ (owing to (3.11)). Assuming that the functions we are using have the regularity requested by the following computations and evaluating (5.4) and (5.2) for $t = 0$, we find

$$q_0(x) = q(x,0) = Q_0\frac{1}{1+l_0(x)}, \quad 0 \le x \le 1, \tag{5.5}$$

where Q_0 is the known constant

$$Q_0 = \frac{p_0(0)}{\displaystyle\int_0^1 \frac{1}{1+l_0(\xi)}\frac{1}{K(b_0(\xi), m_0(\xi), \varepsilon_0(\xi))}d\xi}.$$

Owing to (2.5), (2.16), and (3.16), we have

$$[[\varepsilon]]\,\dot{s} = ((\eta_{(p)} + \theta_{(p)})|_{(s(t)^-,t)} - \Theta)\dot{s} = (ql)|_{(s(t)^-,t)}\,. \tag{5.6}$$

Hence, from (2.17) we get

$$q_c(t) = q(s(t)^-,t)\left(1 + l(s(t)^-,t)\right). \tag{5.7}$$

By combining (5.2) with (5.4) and by taking into account (5.7), we get the following expression for $f(t)$:

$$f(t) = q(0,t) = \frac{p_0(t)}{\displaystyle\int_0^{s(t)} \frac{1}{1+l(\xi,t)}\frac{1}{K(\xi,t)}d\xi + \frac{1-s(t)}{K_0}}, \tag{5.8}$$

where we abridged $K(b(\xi,t), m(\xi,t), \varepsilon(\xi,t))$ by $K(\xi,t)$. Note that both $K(\xi,t)$ and $l(\xi,t)$ are still unknown.

5.2. The fixed point procedure. For the sake of simplicity, we will discuss the case $\alpha_i = 1$, $i = 1,\dots,k$, observing that the general case $\alpha_i < 1$ can be treated with slight modifications.

We start with the following remark. Assume that a pair $(q(x,t), s(t))$ is known to solve the problem. Then, all the other quantities can be computed. Indeed, the species $b_i(x,t)$ are calculated by integrating (3.4) and (3.5); then, $\varepsilon(x,t)$ can be found by means of (3.6). Moreover, the functions m_i, $i = 1,\dots,n$ are computed by means of (3.1) and (3.2). (We will discuss this point with more details in subsection 5.4.) Finally, $p(x,t)$ is found by integrating (3.3).

This argument can be summarized in the following points:

- fix a pair (q,s) in a suitable space;
- solve the problem (3.1)–(3.6), (3.8)–(3.15), (3.17), and (3.18) for (b,m,ε,p) with special adjustments for (3.1) and (3.2), as it will be discussed in detail in subsection 5.4;
- define a new guess $(\tilde{q},\tilde{s})$ in such a way that if it happens that $\tilde{q} = q$, $\tilde{s} = s$, then (3.7) and (3.16) are satisfied.

Such a procedure is performed in the next subsections.

5.3. Auxiliary estimates. Let us consider the set

$$
\begin{aligned}
&\mathcal{E}_T(u_1, u_2, A_y, A_t, M_y, M_t, s_0, A_s, M_s)\\
&= \big\{(u(y,t), s(t)) \mid u \in C^{1,1}(\bar{B}_T),\\
&u(y,0) = q_0(y),\ 0 \le y \le 1, 0 < u_1 \le u(y,t) \le u_2,\\
&\left|\frac{\partial u}{\partial y}(y,t)\right| \le A_y,\ \left|\frac{\partial u}{\partial t}(y,t)\right| \le A_t,\ (y,t) \in \bar{B}_T,\\
&\left|\frac{\partial}{\partial y}u(y_1,t) - \frac{\partial}{\partial y}u(y_2,t)\right| \le M_y\,|y_1 - y_2|\ \forall\, y_1, y_2 \in [0,1],\\
&\left|\frac{\partial}{\partial t}u(y_1,t) - \frac{\partial}{\partial t}u(y_2,t)\right| \le M_t\,|y_1 - y_2|\ \forall\, y_1, y_2 \in [0,1],\\
&s \in C^1[0,T], s(0) = 1, \quad 0 < s_0 \le s(t) \le 1,\\
&-A_s \le \dot{s}(t) \le 0, \quad 0 \le t \le 1,\\
&|\dot{s}(t_1) - \dot{s}(t_2)| \le M_s\,|t_1 - t_2| \quad \forall\, t_1, t_2 \in [0,T]\big\},
\end{aligned}
\tag{5.9}
$$

where $B_T = (0,1) \times (0,T)$ and $u_1 > 0$, $u_2 > u_1$, A_y , A_t, M_y, M_t, s_0, A_s, and M_s are positive constant values to be specified later.

For a pair $(u,s) \in \mathcal{E}_T$, we set

$$
q(x,t) = u\left(\frac{x}{s(t)}, t\right), \quad (x,t) \in \bar{D}_T.
\tag{5.10}
$$

Once problem (3.1)–(3.6), (3.8)–(3.15), (3.17), and (3.18) is solved, we can introduce the mapping $\mathcal{F}$ on $\mathcal{E}_T$ defined as $\mathcal{F}(u,s) = (\tilde{u}, \tilde{s})$, where

$$
\tilde{u}(y,t) = f(t)\frac{1}{1 + l(s(t)y, t)}, \quad (y,t) \in B_T,
\tag{5.11}
$$

$$
\tilde{s}(t) = 1 - \int_0^t \frac{l(s(\tau),\tau)q(s(\tau),\tau)}{\Theta - \big(\eta_{(p)}(s(\tau),\tau) + \theta_{(p)}(s(\tau),\tau)\big)}\,d\tau, \quad t \in [0,T]
\tag{5.12}
$$

inspired by (5.4) (i.e., (3.7)) and (3.16), respectively. In (5.11) the function $f(t)$ is calculated by using (5.8) and $l = \eta_{(p)}/\varepsilon$ (cfr. (5.3)).

Once $(\tilde{u}, \tilde{s})$ is calculated, the new guess $\tilde{q}$ for problem (3.1)–(3.6), (3.8)–(3.15), (3.17), and (3.18) is $\tilde{q}(x,t) = \tilde{u}(x/\tilde{s}(t), t)$ defined in the domain $\{(x,t) : x \in [0, \tilde{s}(t)],\ t \in [0,T]\}$. It is easily checked that if $(\bar{u}, \bar{s})$ is a fixed point of $\mathcal{F}$ in $\mathcal{E}_T$, then $\bar{q} = \bar{u}(x/\bar{s}, t)$ and $\bar{s}$ fulfills (3.7) and (3.16).

PROPOSITION 5.1. *Let* $(u,s) \in \mathcal{E}_T$, $A_y < s_0\varepsilon_0^m/T$ (*see* (4.8)), *with* T *arbitrarily fixed, and set*

$$
\begin{cases}
\epsilon_1 = \varepsilon_0^m - A_y T/s_0,\\
\epsilon_2 = \varepsilon_0^M + \varepsilon_0^m + \|\theta_0^{(k)}\|,\\
C_1 = \|\eta_{(k),0}\|/\varepsilon_0^m + \|\hat{\Phi}\| u_2/\epsilon_1, \quad K_{min} = \min\{K_0, K_m\}
\end{cases}
\tag{5.13}
$$

(see (3.20) *for the definition of* $\hat{\Phi}$*). Then, for* $(x,t) \in D_T$,

$$
0 < \epsilon_1 < \varepsilon(x,t) < \epsilon_2,
\tag{5.14}
$$

$$
0 \le l(x,t) \le C_1,
\tag{5.15}
$$

$$
0 < K_{min} p_0^m \le f(t) \le K_M p_0^M (1 + C_1).
\tag{5.16}
$$

Proof. Inequalities (5.14) follow immediately from the hypotheses and from (3.6), (4.10), and (3.21). As to (5.15), we notice that the function l defined by (5.3) satisfies the following equation in D_T:

$$\begin{cases} \varepsilon \dfrac{\partial l}{\partial t} + q \dfrac{\partial l}{\partial x} = l \dfrac{\partial \theta_{(s)}}{\partial t} - \dfrac{\partial \theta_{(p)}}{\partial t}, \\ l(x,0) = l_0(x) = \dfrac{\eta_{(k),0}(x)}{\varepsilon_0(x)}, \quad l(0,t) = 0. \end{cases} \tag{5.17}$$

Equation (5.17) is obtained by differentiating formally $\eta_{(p)}/\varepsilon$ and by taking into account (3.6) and

$$\frac{\partial \eta_{(p)}}{\partial t} = -\frac{\partial (lq)}{\partial x} - \frac{\partial \theta_{(p)}}{\partial t}, \tag{5.18}$$

which is the sum of (3.1) with respect to i, $i = 1, \dots, k$. The initial and boundary conditions in (5.17) come from (3.8), (3.9), and (3.11).

If we denote by Γ any of the characteristic curves of (5.17), it is easy to see that the slope $d\Gamma(t)\ /dt = q(\Gamma(t),t)\ /\varepsilon(\Gamma(t),t)$ is uniformly bounded by

$$0 < \frac{u_1}{\epsilon_2} \le \frac{d}{dt}\Gamma(t) \le \frac{u_2}{\epsilon_1}. \tag{5.19}$$

The starting points in D_T for the curves Γ are either the points $(x_0, 0)$, with $x_0 \in (0,1)$, and the points $(0, t_0)$, with $t_0 \ge 0$. Correspondingly, the initial value $l(\Gamma(t_0), t_0)$ is $l_0(x)$, $x \in [0,1)$, if $t_0 = 0$ or 0 if $t_0 > 0$. The estimate (5.15) is a consequence of (4.9), (4.10), and (5.19). Finally, the bounds (5.16) for f (see (5.8)) follow from (4.1), (4.2), (4.9), (4.10), and (5.15). □

PROPOSITION 5.2. *If $(u,s) \in \mathcal{E}_T$, then the functions l, $\partial l/\partial x$, and $\partial l/\partial t$ are Lipschitz continuous with respect to x in $\bar{D}_T$.*

Proof. By virtue of the Gronwall's lemma (cf., e.g., [3, Lemma 8.4.1]) and of assumption (4.17), we get

$$\|\theta(x',t) - \theta(x'',t)\|_T \le L_\theta |x' - x''|, \quad L_\theta = L_1 e^{L_2 T} \tag{5.20}$$

with $L_1 = \max_{1\le i\le n} L_{\theta_{i,0}} + \max_{1\le i \le n} L^q_{\hat{S}_i} A_y T/s_0$, $L_2 = \max_{1\le i\le n} L^\theta_{\hat{S}_i}$.

At this point, it is useful to remark that q/ε is Lipschitz continuous with respect to x with constant

$$L_{q/\varepsilon} = \frac{1}{\epsilon_1^2}\left(\epsilon_2 \frac{A_y}{s_0} + u_2 \left(L_{\varepsilon_0 + \theta_0^{(k)}} + \frac{1}{s_0^2} M_y T\right)\right). \tag{5.21}$$

In order to prove the Lipschitz continuity of l, we consider the two characteristic curves Γ_1 and Γ_2 passing through (x',t) and (x'',t), respectively. Recalling (5.20), we easily find

$$|l(x',t) - l(x'',t)| \le \omega\lambda |x' - x''|, \quad t_0 \le \tau \le t, \tag{5.22}$$

where t_0 is the max between the initial times of Γ_1 and Γ_2,

$$\omega = e^{L_{q/\varepsilon} T}, \quad \lambda = \left(T\left(\frac{A_y}{s_0}\tilde{L}_1 + \tilde{L}_2 L_\theta\right) + L_{l_0}\right), \tag{5.23}$$

and $\tilde{L}_1$, $\tilde{L}_2$ depend only on $\|\Phi\|$, $\max_{0\le x\le 1} l_0(x)$, $L^q_{\hat{H}_i} + \sum_{i=1}^k$, $L^q_{\hat{\Phi}_i}$, $L^\theta_{\hat{H}_i} + \sum_{i=1}^k$, $L^\theta_{\hat{\Phi}_i}$ (cf. also (4.17)).

It is worth noticing that (5.22) with (5.17) provides the following bounds for the derivatives of l:

$$\left|\frac{\partial l}{\partial x}\right| \le \omega\lambda, \qquad \left|\frac{\partial l}{\partial t}\right| \le \frac{1}{\epsilon_1}\left(u_2\omega\lambda + \max\{C_1\|\hat{\Phi}\|, \|\hat{H}\|\}\right). \tag{5.24}$$

The Lipschitz continuity of $\partial l/\partial t$, with respect to x can be established arguing in the same way as before. We first remark that $l_t = \partial l/\partial t$ satisfies the following equation along the characteristic curves:

$$\begin{aligned}&\varepsilon(\Gamma(t),t)\frac{\partial}{\partial t}l_t(\Gamma(t),t) + q(\Gamma(t),t)\frac{\partial}{\partial x}l_t(\Gamma(t),t)\\ &= w_1(\Gamma(t),t)l_t(\Gamma(t),t) + w_2(\Gamma(t),t),\end{aligned} \tag{5.25}$$

where

$$w_1 = -\frac{\partial\varepsilon}{\partial t} - \frac{\varepsilon}{q}\frac{\partial q}{\partial t} \quad w_2 = \frac{1}{q}\frac{\partial q}{\partial t}(l\hat{H} - \hat{\Phi}) - \left(l\frac{\partial\hat{H}}{\partial t} - \frac{\partial\hat{\Phi}}{\partial t}\right). \tag{5.26}$$

It can be seen that the Lipschitz constant L_{l_t} is such that

$$L_{l_t} \le \omega\left\{L_{l_t(x,0)} + L_{w_2})e^{\bar{w}_1 T} + (\bar{l}_{t,0} + T\bar{w}_2)L_{w_1}T\right\}, \tag{5.27}$$

where $\bar{w}_i = \max_{(x,t)\in\bar{D}_T}|w_i(x,t)|$, $\bar{l}_{t,0} = \max_{0\le x\le 1}|l_t(x,0)|$. (Note that $l_t(x,0)$ is a known function obtained from (5.17), (3.8), (3.9), (3.4), and (3.5).) It is not difficult to realize that $\bar{w}_i$, $i = 1,2$, does not depend on M_y, M_t.

In order to stress the dependence of $L_{w_i}, i = 1,2$ on M_y, M_t we write the following formula, which can be obtained after some calculations by recalling (3.6) and (5.21):

$$\begin{cases}L_{w_1} \le \alpha_1 M_y + \beta_1 M_t + \gamma_1,\\ L_{w_2} \le \alpha_2 M_y + \beta_2 M_t + \gamma_2 + \delta_2\omega.\end{cases} \tag{5.28}$$

Although we avoid writing, for the sake of simplicity, the lengthy expressions of the coefficients α_i, β_i, γ_i, $i = 1,2,3$, and δ_2, it is important to remark that they depend only on A_y, A_t, A_s, s_0, u_1, u_2, T, L_{l_0}, $L_{l_t(x,0)}$ and on the Lipschitz constants appearing in (4.17).

Owing to (5.28), we finally write

$$L_{l_t} \le \omega(\alpha_3 M_y + \beta_3 M_y + \gamma_3 + \delta_3\omega) \tag{5.29}$$

with

$$\begin{cases}\alpha_3 = \alpha_1 T(L_{l_0} + T\bar{w}_2) + \alpha_2 T e^{\bar{w}_1 T},\\ \beta_3 = \beta_1 T(L_{l_0} + T\bar{w}_2) + \beta_2 T e^{\bar{w}_1 T},\\ \gamma_3 = L_{l_t(x,0)}e^{\bar{w}_1 T} + \gamma_1 T(L_{l_0} + T\bar{w}_2) + \gamma_2 T e^{\bar{w}_1 T},\\ \delta_3 = \delta_2 T e^{\bar{w}_1 T}.\end{cases} \tag{5.30}$$

Note that the values defined in (5.30) do not depend on M_y, M_t.

We conclude the proof of this lemma by remarking that the Lipschitz continuity of $\partial l/\partial x$ with respect to x comes directly from (5.17):

$$L_{l_x} \leq \left(\frac{\epsilon_1}{u_1}\right)^2 \max_{(x,t)\in\bar{D}_t} |l_t(x,t)| + \frac{\epsilon_2}{u_1} L_{l_t} + \lambda \frac{\|\hat{H}\|}{u_1}\omega + \frac{1}{u_1}\left\{C_1 L_{\theta_{(s)_t}} + L_{\theta_{(p)_t}} + \frac{A_y}{s_0 u_1}\left(C_1\|\hat{H}\| + \|\hat{\Phi}\|\right)\right\}, \tag{5.31}$$

where $|l_t|$ is estimated by means of (5.24). □

5.4. Determination of the concentrations. In order to calculate $f(t)$ by means of (5.8), it is necessary to find the concentration of each species separately, because of the dependence of f on θ_i, η_i, $i = 1, \ldots, n$, through the hydraulic conductivity K.

Problem (3.1), (3.8), and (3.11) involves the x-derivative of q/ε, which does not necessarily exist, since (see (3.6)) ε may not be differentiable with respect to x. Hence, we formally consider the following equation, which comes from (3.7) and (5.4) and is of course applicable to the solution of the original problem only:

$$\frac{\partial \varepsilon}{\partial x} = \varepsilon\left(\frac{1}{q}\frac{\partial q}{\partial x} + \frac{1}{\varepsilon + \eta_{(p)}}\frac{\partial}{\partial x}(\varepsilon + \eta_{(p)})\right). \tag{5.32}$$

If we recall (2.5), we can replace the x-derivative of $\varepsilon + \eta_{(p)}$ on the right-hand side by $-\theta_0'(x)$ $-\partial/\partial x(\theta_{(p)}$ $+\theta_{(s)})$, which is computed by introducing the vector

$$\zeta = \frac{\partial}{\partial x}(\theta_1, \ldots, \theta_n)$$

and by solving the ODEs

$$\begin{cases} \dfrac{\partial \zeta_i}{\partial t} = -\dfrac{\partial \hat{S}_i}{\partial q}\dfrac{\partial q}{\partial x} - \nabla_\theta \hat{S}_i \cdot \zeta & \text{for } i = 1, \ldots, n, \\ \zeta(x,0) = \zeta_0(x) = (\theta_1'(x), \ldots, \theta_n'(x)). \end{cases} \tag{5.33}$$

Having in mind (5.32), we modify (3.1) replacing $\partial/\partial x(q/\varepsilon)$ by the x-derivative of a function E such that

$$\frac{\partial E}{\partial x} = \frac{1}{\varepsilon + \eta_{(p)}}\frac{q}{\varepsilon}\left(\zeta_0 + \sum_{i=1}^n \zeta_i\right). \tag{5.34}$$

The Lipschitz continuity of ζ_i (hence that of $\partial E/\partial x$) can be easily checked by examining (5.33). Thus, for the components $i = 1, \ldots, k$ we consider the following modified problem, to be solved in the given domain $D_T = \{(x,t) : 0 < x < s(t),\ 0 < t < T\}$ together with (3.8) and (3.11):

$$\frac{\partial \eta_i}{\partial t} + \alpha_i \eta_i \frac{\partial E}{\partial x} + \alpha_i \frac{q}{\varepsilon}\frac{\partial \eta_i}{\partial x} = -\frac{\partial \theta_i}{\partial t}, \quad i = 1, \ldots, k. \tag{5.35}$$

As for the remaining species $i = k+1, \ldots, n$, we define

$$c_i = \frac{m_i}{\varepsilon}, \quad i = k+1, \ldots, n,$$

as the concentration of the i-component in the solute. Owing to (3.2) and (3.6), we see that in the original problem c_i must satisfy the following equation in $D_T \cup R_T$:

$$\begin{aligned}\varepsilon\frac{\partial c_i}{\partial t} + \frac{\partial}{\partial x}\left(-\varepsilon D_i \frac{\partial c_i}{\partial x}\right) + q\frac{\partial c_i}{\partial x}\\ -c_i\frac{\partial \theta_{(s)}}{\partial t} = H_i,\ i = k+1, \dots, n.\end{aligned} \tag{5.36}$$

Actually, here too we use a modified equation, replacing $\partial\varepsilon/\partial x$ with

$$\frac{\varepsilon}{q}\left(\frac{\partial q}{\partial x} - \varepsilon\frac{\partial E}{\partial x}\right).$$

The modified equation replacing (5.36) is therefore

$$\begin{aligned}\varepsilon\frac{\partial c_i}{\partial t} + \varepsilon\frac{\partial}{\partial x}\left(-D_i\frac{\partial c_i}{\partial x}\right) + \left(q - D_i\frac{\varepsilon}{q}\left(\frac{\partial q}{\partial x} - \varepsilon\frac{\partial E}{\partial x}\right)\right)\frac{\partial c_i}{\partial x}\\ -c_i\frac{\partial \theta_{(s)}}{\partial t} = H_i, \quad i = k+1, \dots, n.\end{aligned} \tag{5.37}$$

The initial and boundary conditions associated to (5.36) come from (3.8), (3.9), (3.14), (3.15), and (3.17), namely,

$$c_i(x,0) = m_{i,0}(x)/\varepsilon_0(x), \quad i = k+1, \dots, n, \tag{5.38}$$

$$[[c_i]] = 0, \quad x = s(t) \quad i = k+1, \dots, n, \tag{5.39}$$

$$\left[\left[\varepsilon D_i\frac{\partial c_i}{\partial x}\right]\right] = 0 \quad x = s(t), \quad i = k+1, \dots, n, \tag{5.40}$$

$$\frac{\partial c_i}{\partial x}(1,t) = 0, \quad 0 < t < T, \quad i = k+1, \dots, n. \tag{5.41}$$

Existence for the diffraction problem (5.37)–(5.41) is guaranteed by Theorem 13.1, page 227, of [11].

5.5. The property $\mathcal{F}(\mathcal{E}_T) \subseteq \mathcal{E}_T$. In order to simplify the mathematical presentation, it can be assumed that the dependence of K on m_i occurs through q (i.e., $K = K(q,b,\varepsilon)$). Even if we are going to make such an assumption, it must be said that it plays a very marginal role in the proof of the proposition we are going to show.

PROPOSITION 5.3. *There exists at least one set of positive values* $\{T, u_1, u_2 > u_1,\ A_y,\ A_t,\ M_y,\ M_t,\ s_0,\ A_s, M_s\}$ *so that, if* $(u,s) \in \mathcal{E}_T(u_1, u_2,\ A_y, A_t, M_y,\ M_t, s_0, A_s, M_s)$*, then* $(\tilde u, \tilde s) \in \mathcal{E}_T$*, provided that* (*see* (5.13) *for the definition of* C_1)

(i)

$$\epsilon_2 C_1 + \|\theta_{(k),0}\| < \Theta; \tag{5.42}$$

(ii)

$$\frac{\|\hat\Phi\|}{\epsilon_1}K_M p_0^M < 1, \quad p_0^M(1+C_1)^2\left|\frac{\partial K}{\partial q}\right| < 1; \tag{5.43}$$

(iii) *the Lipschitz constants with respect to* x *of the given functions* ε_0*,* $\eta_{(k),0}$*,* $\eta_0^{(k)}$*,* ε_0'*,* $\eta_{(k),0}'$*, and the Lipschitz constants* $L^q_{\hat S_i}$*,* $L^\theta_{\hat S_i}$*,* $i = 1, \dots, n$*, of the removal functions are small enough, in the sense that will be specified more precisely through the proof of the proposition.*

Proof. It is easy to check that $\tilde{u}(y,0) = u(y,0) = q_0(y)$, $0 \leq y \leq 1$, where q_0 is given by (5.5). From (5.11), (5.15), and (5.16) we see that

$$\frac{1}{1+C_1} K_m p_0^m \leq \tilde{u}(x,t) \leq K_M p_0^M (1+C_1). \tag{5.44}$$

Recalling (5.13) and owing to the first inequality in (5.43), it is possible to find u_2 sufficiently large so that $K_M p_0^M\ (1+C_1) \leq u_2$. On the other hand, it is sufficient to take $u_1 = (1+C_1)^{-1}(K_m p_0^m)$ in order to have $\tilde{u}$ bounded within the interval $[u_1, u_2]$. From (5.15), (5.14), and (4.10) we see that

$$1 \geq \tilde{s} \geq 1 - \frac{K_M p_0^M C_1 (1+C_1) T}{\Theta - (\epsilon_2 C_1 + \sup\limits_{0 \leq x \leq 1} \theta_{(k),0}(x))}. \tag{5.45}$$

By virtue of (5.42) and taking T sufficiently small, we see that the boundary $\tilde{s}$ is uniformly bounded by a constant $s_0 > 0$.

If we calculate explicitly the y-derivative of $\tilde{u}$ and we recall (5.16) and (5.24), we can easily realize that

$$\left|\frac{\partial \tilde{u}}{\partial y}\right| \leq (1+C_1) K_M p_0^M \omega \lambda. \tag{5.46}$$

Recalling (5.23), it can be seen that two values A_{y_1} and A_{y_2}, $0 < A_{y_1} < A_{y_2}$ can be found such that the right-hand side of (5.46) is smaller than A_y for $A_y \in [A_{y_1}, A_{y_2}]$, provided that $\tilde{L}_1$ and $\tilde{L}_2$ are taken sufficiently small. Note that $\tilde{L}_1$ and $\tilde{L}_2$ depend only on the quantities mentioned in assumption (iii) of the statement of the present proposition. It is useful to observe that A_{y_1} tends to zero if $\tilde{L}_1$ and $\tilde{L}_2$ tend to vanish.

As to the t-derivative of $\tilde{u}$, we have from (5.24) and (5.46)

$$\left|\frac{\partial \tilde{u}}{\partial t}\right| \leq \max_{0 \leq t \leq T} |\dot{f}(t)| + (1+C_1) K_M p_0^M \left(\omega\lambda \left(A_s + \frac{u_2}{\epsilon_1} \right) + \frac{1}{\epsilon_1} \max\{C_1 \|\hat{\Phi}\|, \|\hat{H}\|\} \right). \tag{5.47}$$

Recalling (5.13) and (5.22), we see that the second term on the right-hand side of (5.47) is independent of A_t. On the other hand, when we calculate $\dot{f}$ explicitly, it is easy to find a bound for it of the type $c_1 A_t + c_2$, where c_1 is exactly the left-hand side of (5.43), the second inequality, and c_2 is independent of A_t. Hence, $c_1 < 1$ entails that $|\partial\tilde{u}/\partial t| \leq A_t$ for $A_t > c_2(1-c_1)^{-1}$.

We now pass evaluating the Lipschitz constants $L_{\tilde{u}_y}$ and $L_{\tilde{u}_t}$ with respect to y of the derivatives of $\tilde{u}$. We find (cf. (5.22), (5.27), (5.31), and (5.24))

$$L_{\tilde{u}_y} \leq K_M p_0^M (1+C_1)^2 \left((1+C_1) L_{l_x} + 2\omega^2\lambda^2 \right), \tag{5.48}$$

$$\begin{aligned} L_{\tilde{u}_t} \leq \max_{0 \leq t \leq T} \|\dot{f}\| + K_M p_0^M (1+C_1)^2 [(1+C_1)(L_{l_x} A_s + L_{l_t}) \\ + 2\omega\lambda \left(\omega\lambda A_s + \frac{1}{\epsilon_1}(u_2 \omega\lambda + C_1 \|\hat{\Phi}\| + \|\hat{H}\|) \right)]. \end{aligned} \tag{5.49}$$

We first remark that $\max_{0 \leq t \leq T} |\dot{f}|$ can be expressed in the form $\gamma_f + \delta_f \omega$, where γ_f and δ_f depend also on the norm of $|\partial K/\partial q|$, $|\partial K/\partial \varepsilon|$, $|\dot{p}_0|$, but not on M_y, M_t.

Imposing that the right-hand sides of (5.48) and (5.49) are not greater than M_y and M_t, respectively, and taking account of the estimates (5.29) and (5.31), we arrive at the following system of inequalities:

$$\begin{cases} (\mu_1\omega+\mu_2)M_y+\mu_3\omega M_t+\mu_4\omega^2+\mu_5\omega\le M_y,\\ A_s\left((\mu_1\omega+\mu_2)M_y+\mu_3\omega M_t\right)+\mu_6\omega^2+(\mu_7+\delta_f)\omega+\gamma_f\le M_t, \end{cases} \tag{5.50}$$

where (see also (5.30))

$$\begin{cases} \mu_1=(1+C_1)\nu_0\left(\dfrac{\epsilon_2}{u_1}\alpha_3+\dfrac{\epsilon_1 r_2}{u_1^2}\lambda\right),\\ \mu_2=(1+C_1)\nu_0\dfrac{1}{\epsilon_1u_1^2s_0^2}(C_1\|\hat{\Phi}\|+\|\hat{H}\|)T,\\ \mu_3=\nu_0\dfrac{\epsilon_2}{u_1}\beta_3,\\ \mu_4=\nu_0\left\{\left[(1+C_1)\dfrac{\epsilon_2}{u_1}\delta_3\right]+2\dfrac{\epsilon_2}{\epsilon_1u_2}\lambda^2\right\},\\ \mu_5=\dfrac{\nu_0}{u_1}\left\{(1+C_1)\left[\gamma_3\epsilon_2u_1+\lambda\|\hat{H}\|+\dfrac{u_2\epsilon_1r_1\lambda}{u_1}\right]\right.\\ \qquad\qquad\left.+2\lambda\left(\left(1+\dfrac{\epsilon_2}{\epsilon_1}\right)(C_1\|\hat{\Phi}\|+\|\hat{H}\|)\right)\right\} \end{cases} \tag{5.51}$$

and

$$\begin{cases} \nu_0=K_Mp_0^M(1+C_1)^2,\\ r_1=\dfrac{1}{\epsilon_1^2}\left(\epsilon_2\dfrac{A_y}{s_0}+u_2L_{\varepsilon_0+\theta_0^{(k)}}\right)T,\\ r_2=\dfrac{u_2}{\epsilon_1^2s_0^2}T^2. \end{cases} \tag{5.52}$$

The coefficient μ_6 (resp., μ_7) is similar to μ_4 (resp., μ_5), with the only difference being that A_s multiplies the term in square brackets. We recall that ω depends on M_y through $L_{q/\varepsilon}$ (cf. (5.21) and (5.23)), namely,

$$\omega=e^{r_1+r_2M_y}.$$

Now, fix any $Y>1$. Owing to (5.51) and (5.52) (see also (5.30) for the definition of α_3), we easily realize that a positive value T_0 can be found such that

$$e^{r_1(T)}<Y<\frac{1-\mu_1(T)}{\mu_2(T)} \tag{5.53}$$

for any $T\in(0,T_0]$. Second, we define

$$Z=3\left(\delta_fe^{r_1(T_0)+r_2(T_0)Y}+\gamma_f\right) \tag{5.54}$$

(we recall that γ_f and δ_f have been introduced just above (5.49)) and we pass to solve the first inequality in (5.50) setting $M_t=Z$. We find it convenient to write such an inequality in the following form:

$$\begin{cases} \phi_1(\omega,T) \ge \phi_2(\omega,T), \\ \phi_1(\omega,T) = \dfrac{\ln\omega - r_1}{r_2}, \\ \phi_2(\omega,T) = \dfrac{(\mu_3 M_t + \mu_5)\omega + \mu_4\omega^2}{1-\mu_1\omega-\mu_2}. \end{cases} \tag{5.55}$$

The dependence of ϕ_1 and ϕ_2 on T is through the coefficients μ_i, $i=1,\dots,5$, and r_1, r_2. It can be seen, by simply examining the functions ϕ_1 and ϕ_2 and the coefficients defined in (5.51) and (5.52), that two positive values $T_1 \le T_0$ and $\omega_0 > e^{r_1(T_0)}$ can be found such that (5.55) is satisfied for $0 < T \le T_1$ and $\omega_0 \le \omega \le Y$.

On the other hand, we rewrite the second inequality in (5.50) as

$$\begin{cases} Z \ge \phi_3(\omega,T) + \phi_4(\omega,T), \\ \phi_3(\omega,T) = \dfrac{A_s(\mu_1\omega+\mu_2)\phi_1(\omega,T) + \mu_6\omega^2 + \mu_7\omega}{1 - A_s\mu_3\omega}, \\ \phi_4(\omega,T) = \dfrac{\delta_f\omega + \gamma_f}{1 - A_s\mu_3\omega}. \end{cases} \tag{5.56}$$

Note that $\phi_3(\omega,T) \le \phi_3(Y,T)$ for $\omega \in [\omega_0, Y]$. Recalling also (5.30) and (5.51), we see that it is possible to find a positive $T_2 \le T_1$ so that

$$1 - A_s\mu_3 \ge \frac{1}{2}, \quad \phi_3(Y,T) \le \frac{1}{3}Z.$$

Since $\phi_4(\omega,T) \le \frac{2}{3}Z$ (see (5.54)) for $T \in (0,T_2]$ and $\omega \in [\omega_0, Y]$, we conclude that (5.56) (hence (5.50) with $M_t = Z$) holds for T and ω belonging to the same intervals.

By definition (5.12) and estimate (5.15) we immediately see that the boundary velocity $\dot{\tilde{s}}$ is uniformly bounded as follows:

$$-\frac{u_2 C_1}{\Theta - (\epsilon_2 C_1 + \|\theta_{(k),0}\|)} \le \dot{\tilde{s}}(t) \le 0. \tag{5.57}$$

Hence, it is sufficient to take $A_s \ge u_2 C_1 \; (\Theta - (\epsilon_2 C_1 + \; \|\theta_{(k),0})\|)^{-1}$ in order to have a uniform bound for $\dot{\tilde{s}}(t)$. Note that $A_s > 0$ because of (5.42).

Finally, we discuss the Lipschitz continuity of $\dot{\tilde{s}}$. We take $0 \le t_1 < t_2$ and we consider the point $(\Gamma_2(t_1), t_1)$, where Γ_2 is the characteristic curve passing by $(\tilde{s}(t_2), t_2)$. Calling

$$\sigma(t) = \frac{l(\Gamma_2(t),t)q(\Gamma_2(t),t)}{\Theta - (\eta_{(p)}(s(t),t) + \theta_{(p)}(s(t),t))},$$

we see that

$$\begin{aligned} &|\dot{\tilde{s}}(t_1) - \dot{\tilde{s}}(t_2)| \le |\dot{\tilde{s}}(t_1) - \sigma(t_1)| + |\sigma(t_1) - \dot{\tilde{s}}(t_2)| \\ &\le \frac{1}{(A_y T)^2}\left((\Theta + C_1 u_2 L_{\eta_{(p)}+\theta_{(p)}})|\Gamma_2(t_1) - \tilde{s}(t_1)|\right. \\ &+\Theta\left|(lq)|_{(\Gamma_2(t_1),t_1)} - (lq)|_{(\Gamma_2(t_1),t_1)}\right| \\ &\left.+C_1 u_2\left|(\eta_{(p)} + \theta_{(p)})|_{(\Gamma_2(t_1),t_1)} - (\eta_{(p)} + \theta_{(p)})|_{(\Gamma_2(t_2),t_2)}\right|\right). \end{aligned} \tag{5.58}$$

By virtue of (5.19) we easily get

$$|\tilde{s}(t_1) - \Gamma_2(t_1)| \leq \left(A_s + \frac{u_2}{\epsilon_1}\right)|t_1 - t_2|. \tag{5.59}$$

On the other hand, the functions q, l, $\eta_{(p)}$, and $\theta_{(p)}$ are easily estimated along a characteristic curve. Eventually, the right-hand side of (5.58) is bounded by a term of the type $C_2|t_1 - t_2|$, with C_2 independent of M_s. □

5.6. The main result. We start by the following local result. Second, it will be extended globally in time (see Proposition 5.5).

PROPOSITION 5.4. *Under the same hypotheses as in Propositions* 5.1 *and* 5.3*, $\mathcal{F}$ has at exactly one fixed point in $\mathcal{E}_T$, for some $T > 0$.*

Proof. Concerning existence, we make use of the Schauder's fixed point theorem. It is easily checked that the set $\mathcal{E}_T$ is bounded, closed, and convex. By virtue of Proposition 5.3, there exists $T > 0$ so that $\mathcal{F}(\mathcal{E}_T) \subseteq \mathcal{E}_T$. As to the continuity of $\mathcal{F}$, we take two points (u_1, s_1), $(u_2, s_2) \in \mathcal{E}_T$ and the two groups of calculated functions $(\varepsilon_1, \theta_1^{(1)}, \ldots, \theta_n^{(1)}, \eta_1^{(1)}, \ldots, \eta_n^{(1)})$, $(\varepsilon_2, \theta_1^{(2)}, \ldots, \theta_n^{(2)}, \eta_1^{(2)}, \ldots, \eta_n^{(2)})$. If $\|u_1 - u_2\|_{C^{1,1}}$ $+\|s_1 - s_2\|_{C^1}$ is sufficiently small, it is not difficult to see that $|\varepsilon_1 - \varepsilon_2|$, $\|\theta^{(1)} - \theta^{(2)}\|_T$, and $\|\eta^{(1)} - \eta^{(2)}\|_T$ are arbitrarily small (of course, we have to restrict x in $[0, \min_{i=1,2} s_i(t)$, for each $t \in [0, T])$. The same property is true for the derivatives of u with respect to y and t and for the t-derivative of s. Eventually, it can be obtained that $\|\tilde{u}_1 - \tilde{u}_2\|_{C^{1,1}}$ $+\|\tilde{s}_1 - \tilde{s}_2\|_{C^1} \to 0$ for $\|u_1 - u_2\|_{C^{1,1}}$ and $\|s_1 - s_2\|_{C^1}$ tending to zero.

Finally, the precompactness of $\mathcal{F}(\mathcal{E}_T)$ follows from what was proved in Propositions 5.2 and 5.3 and from observing that $\partial\tilde{u}/\partial t$ is Lipschitz continuous also with respect to t. Indeed, for any pair (y, t_1), (y, t_2), $0 \leq t_1 < t_2$, $0 \leq y \leq 1$, we consider the points $P_1 \equiv (s(t_1)y, t_1)$, $P_2 \equiv (s(t_2)y, t_2)$ and $P_3 \equiv (\Gamma_2(t_1), t_1)$ with Γ_2 characteristic curve passing by P_2. By examining (5.25), it can be seen that $l_t(\Gamma_2(t), t)$ (hence l_x) is Lipschitz continuous with respect to t along the characteristic $\widehat{P_2P_3}$. On the other hand, the Lipschitz continuity (with respect to y) along the horizontal segment $\overline{P_1P_3}$ has been already proved. Moreover, for any $y \in [0, 1]$, $0 \leq t_1 < t_2$ we have (cf. (5.19))

$$|y - y'| \leq \frac{u_2}{s_0\epsilon_1}|t_1 - t_2|, \quad y' = \Gamma_2(t_1).$$

As to uniqueness, we observe that the mapping is a contraction if T is chosen sufficiently small. By means of an iteration argument, we can extend uniqueness with no limitation in time. This concludes the proof of Proposition 5.4. □

We now define

$$\beta = \frac{\|\hat{\Phi}\|K_M p_0^M}{\varepsilon_0^m}, \quad \epsilon_2 = \varepsilon_0^M + \varepsilon_0^m + \|\theta_0^{(k)}\|, \tag{5.60}$$

where the norm is the sup-norm and we take $\bar{C}_1$ as any constant such that

$$\frac{\dfrac{\|\eta_{(p),0}\|}{\varepsilon_0^m} + \beta}{1 - \beta} < C_1 < \frac{\|\theta_{(p,0)}\|}{\epsilon_2}.$$

PROPOSITION 5.5. *Under assumptions*

(i) *a* $\beta < 1$, $\frac{\epsilon_2/\varepsilon_0^m}{1-\beta}\|\eta_{(p),0}\| + \|\theta_{(k),0}\| + \epsilon_2\frac{1}{1-\beta} < \Theta$,

(ii) *a* $p_0^M(1+\bar{C}_1)^2|\frac{\partial K}{\partial q}| < 1$,

and assumption (iii) *of Proposition* 5.3*, system* (3.1)–(3.19) *has exactly one solution for any* $t \geq 0$.

Proof. It is immediate to see that if (i) part a and (ii) part a hold, then a positive T can be found such that assumptions (i) and (ii) of Proposition 5.3 are fulfilled.

We also remark the following important property of the solution found by means of Proposition 5.4. By integrating (3.1) in the domain D_t, $0 < t \leq T$, by using the Gauss–Green formula, and by taking into account of the boundary condition (3.16) we get the following balance:

$$\begin{aligned}(5.61)\qquad &\int_0^{s(t)} (\eta_{(p)}(x,t) + \theta_{(p)}(x,t))dx + \Theta(1-s(t))\\ &= \int_0^1 (\eta_{(k),0}(x) + \theta_{(k),0}(x))dx.\end{aligned}$$

The physical meaning of (5.61) is evident: the initial concentration of the i-species, $i = 1, \dots, k$, is partitioned between the porous medium and the compact layer. When the first term in the left-hand side of (5.61) vanishes, we get the minimum possible value s_{min} for the boundary $s(t)$:

$$(5.62)\qquad s_{min} = 1 - \frac{\bar{m}_0}{\Theta}, \quad \bar{m}_0 = \int_0^1 (\eta_{(k),0} + \theta_{(k),0})dx.$$

In (5.62) $\bar{m}_0$ is the initial volume occupied by the particles. Note that $0 < s_{min} < 1$, by virtue of (4.3).

An important consequence of (5.61) is that s is bounded by a quantity depending only on the data of the problem and independently of T. A second property of the solution is

$$(5.63)\qquad \eta_{(p)} + \theta_{(p)} < \Theta, \quad (x,t) \in D_T.$$

Inequality (5.63) follows immediately from (5.15), (5.14), (4.9), and (5.42) and will be commented in the following section.

Assume now that $\eta_{(p)}(x,T) \not\equiv 0$, $x \in [0, s(T)]$. (Otherwise, the removal process stops, as will be stated more precisely in subsection 6.1.) Using (2.5) (note that this equation is a consequence of (3.1), (3.6), (3.7), and (3.22)) and taking into account (5.63) and the fact that $\theta_{(p)}$ and $\theta_{(s)}$ are decreasing functions with respect to t, we can estimate the porosity at the time $t = T$ as follows:

$$\begin{aligned}(5.64)\qquad & 1 - (\Theta + \theta_0(x) + {\theta_{(s)}}_0(x)) < \varepsilon(x,T)\\ (5.65)\qquad & < 1 - \theta_0(x), \quad x \in [0, s(T)].\end{aligned}$$

We remark that $s(T) > 0$ owing to (5.62). Recalling (4.7) and (4.8), we see that

$$(5.66)\qquad \varepsilon_0^m < \varepsilon(x,T) < \varepsilon_0^M, \quad x \in [0, s(T)].$$

Let us consider now the interval $[T, 2T]$. Arguing as in Proposition 5.1, we see that $l(x,t)$ is bounded by the same constant C_1 for $T \leq t \leq 2T$. (Note that ϵ_1 and ϵ_2 have the same value as for $t \in [0,T]$, by virtue of (5.66).) This in turn implies that assumptions (i)–(iii) are still valid for $t = T$, so that the result stated in Proposition 5.4 still holds for $t \in [T, 2T]$. Iterating the procedure, we get existence and uniqueness globally in time. □

6. Qualitative properties of the solution.

Let us define

$$D = \{(x,t) : 0 < x < s(t), t > 0\}.$$

We start by commenting (5.63), which prevents the formation of new compact layers in the region D. First, we remark that condition (5.42), which is fundamental in order to get (5.63), is more restrictive than the first inequality in (4.4), since it can be written in the following form (cf. (5.13)):

$$\left(1 + \frac{\varepsilon_0^M + \|\theta_0^{(k)}\|}{\varepsilon_0^m}\right) \|\eta_{(k),0}\| + \|\theta_{(k),0}\| \tag{6.1}$$

$$+ \left(\varepsilon_0^m + \varepsilon_0^M + \|\eta_{(s)_0}\|\right) \|\hat{\Phi}\| \frac{u_2}{\epsilon_1} < \Theta. \tag{6.2}$$

It is worth noticing that (6.1) is a condition either on the initial distribution of particles $\theta_{(k),0}$, $\eta_{(k),0}$ or on the removal rate of the particles $\|\hat{\Phi}\|$.

If a uniform distribution of m_i, b_i, $i = 1, \ldots, k$, is assumed for $t = 0$ (as in [6, part 1]), where the case $k = 1$ is discussed), a condition like $\eta_{(k),0} + \theta_{(k),0} < \Theta$ is sufficient in order to guarantee (5.63) for $t > 0$. However, in the present case it is not difficult to exhibit nonuniform initial distributions of m_i and b_i (even if $k = 1$) such that $\eta_{(p)} + \theta_{(p)} = \Theta$ for some point in D. In such a case, a new compact layer will develop.

6.1. Growth of the compact layer.

We start by remarking that, by virtue of (4.8), (4.2), and (5.5), we have that $\dot{s}(t) = 0$ if and only if $\eta_{(p)}(s(t), t) = 0$.

We now introduce the set

$$\mathcal{S} = \{(x,t) \in D \mid b_i(x,t) \le \beta_i(q,b),\ i = 1, \ldots, k\}, \tag{6.3}$$

where β_i are the functions appearing in (4.9). Furthermore, we set $\mathcal{R} = D/\mathcal{S}$.

Proposition 6.1. *If a characteristic curve* Γ *with* $t_0 > 0$ *is such that* $\Gamma \subset \mathcal{S}$, *then* $\dot{s}(\tau) = 0$, *where* τ *is the time when the curve* Γ *intersects the boundary* s.

Proof. It is a consequence of the following formula, which comes from integrating along a characteristic curve Γ:

$$\begin{aligned} l(\Gamma(t), t) = {} & l(\Gamma(t_0), t_0) e^{-\int_{t_0}^t \hat{H}(q(\Gamma(\tau),\tau),\theta(\Gamma(\tau),\tau))d\tau} \\ & + \int_{t_0}^t \hat{\Phi}(q(\Gamma(\tau),\tau),\theta(\Gamma(\tau),\tau)) e^{-\int_\tau^t \hat{H}(q(\Gamma(\sigma),\sigma),\theta(\Gamma(\sigma),\sigma))d\sigma} d\tau. \end{aligned} \tag{6.4}$$

The first term in the right-hand side of (6.4) is zero, since $\Gamma(t_0) = 0$; moreover, the second term (due to the removal of particles) vanishes owing to (cfr. (3.20). Note that τ is certainly finite, because of (5.19). ☐

A further consequence of (6.4) is that if $\Gamma \bigcap \mathcal{R} \neq \emptyset$, then $\dot{s}(\tau) < 0$, with τ defined as above. The profile of the characteristic curves and of the free boundary $s(t)$ is sketched in Figure 6.1.

We consider now the region $\mathcal{S}_1 = \bigcup_{\Gamma \in \mathcal{S}} \Gamma \subseteq D$. Since $\eta_{(p)} = 0$, $\theta_{(p)}$ is constant in that region, from Proposition 6.1, (3.7), and (5.18)

$$\frac{\partial q}{\partial x}(x,t) = 0, \quad \dot{s}(t) = 0, \quad (x,t) \in \mathcal{S}_1. \tag{6.5}$$

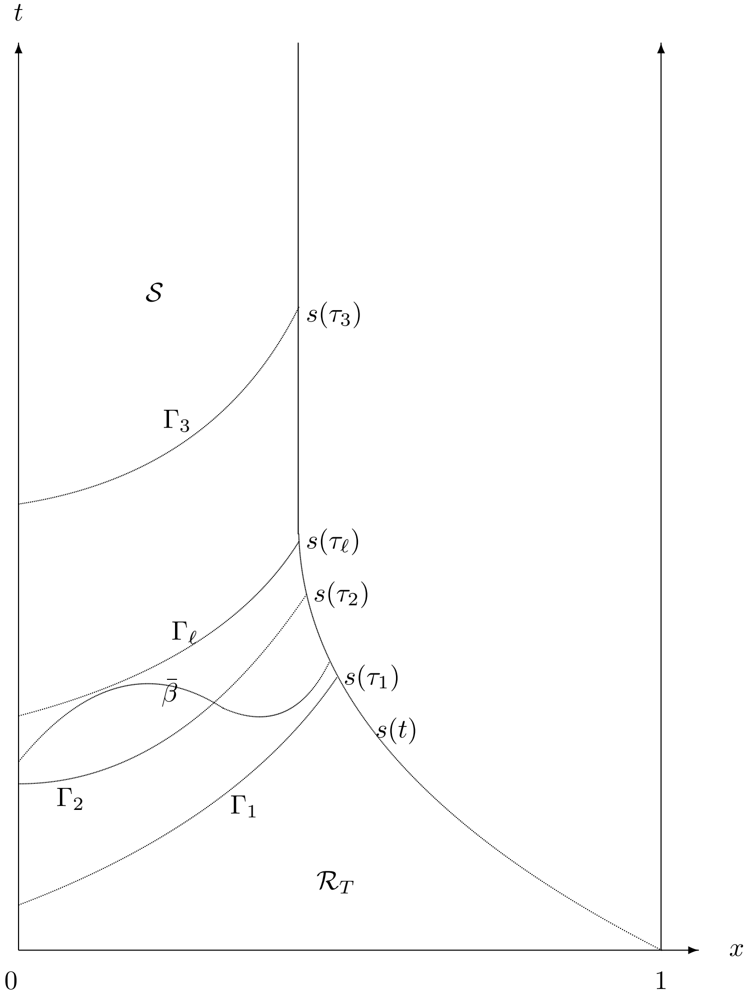


FIG. 6.1. *The characteristic curves Γ_1 and Γ_2 are totally or partially contained in $\mathcal{R}$; hence $\dot{s}(\tau_1)$ and $\dot{s}(\tau_2)$ are negative; Γ_3 is contained in $\mathcal{S}$; hence $\dot{s}(\tau_3) = 0$. The curve $\bar{\beta}$ separates $\mathcal{S}$ from $\mathcal{R}$, while the curve Γ_ℓ is the lower boundary of the region $\mathcal{S}_1$. For $t \geq \tau_\ell$, we have $\dot{s}(t) = 0$.*

The value of s in $\mathcal{S}_1$ is obtained by solving the equation (cf. (5.8))

$$q_c(t)\left(\frac{1-s}{K_0} + \int_0^s \frac{1}{K(\xi,t)}d\xi\right) = p_0(t). \tag{6.6}$$

Notice that $K(\xi, t)$ changes with respect to t in $\mathcal{S}_1$ only if ε varies. Moreover,

$$\frac{\partial K}{\partial t} = -\frac{\partial K}{\partial \varepsilon}\frac{\partial \theta_{(s)}}{\partial t}, \quad (x,t) \in \mathcal{S}_1. \tag{6.7}$$

Equation (6.7) comes from (3.6) and (6.5). Owing to (6.7), (6.6) and recalling (4.10), (4.11) we can conclude that

- if $\partial K/\partial \varepsilon \geq 0$, then the water flux $q_c(t)$ increases (and also q because of (6.5)) if $p_0(t)$ increases and the removal process may restart ($\dot{s} < 0$);

- if $\partial K/\partial\varepsilon \le 0$ and $p_0(t)$ does not increase, then $q_c(t)$ is a nonincreasing function and the removal process cannot occur again.

If we neglect the dependence of K on ε, we have simply that $q_c(t)$ increases if and only if $p_0(t)$ increases; thus, only in that case the removal process can occur a second time.

Owing to (5.62), the limit of $s(t)$ for t tending to ∞ is a positive value not smaller than s_{min}.

REFERENCES

[1] L. Billi, *Incompressible flows through porous media with temperature-dependent parameters*, Nonlinear Anal., 31 (1998), pp. 363–383.

[2] L. Billi, *Non-isothermal flows in porous media with curing*, European J. Appl. Math., 8 (1997), pp. 623–637.

[3] J. R. Cannon, *The One-Dimensional Heat Equation*, Encyclopedia Math. Appl. 23, Addison-Wesley, Reading, MA, 1984.

[4] E. Comparini, P. Mannucci, *Penetration of a wetting front in a porous medium interacting with the flow*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 425–438.

[5] E. L. Cussler, *Diffusion: Mass Transfer in Fluid Systems*, Cambridge University Press, Cambridge, UK, 1984.

[6] A. Fasano, *Some non-standard one-dimensional filtration problems*, The Bulletin of the Faculty of Education, Chiba University, Japan, 44 (1996), pp. 5–29.

[7] A. Fasano, *The penetration of a wetting front through a porous medium accompanied by the dissolution of a substance*, in International Congress on Math Modelling of Flow in Porous Media, A. Bourgeat et al., eds., World Scientific, River Edge, NJ, 1995, pp. 183–195.

[8] A. Fasano and M. Primicerio, *Flows through saturated mass exchanging porous media under high pressure gradients*, in Proceedings of Calculus of Variations, Applications and Computations, C. Bandle et al., eds., Pitman Res. Notes Math. Ser. 326, Longman Sci. Tech., Harlow, UK, 1994, pp. 109–129.

[9] A. Fasano and M. Primicerio, *Mathematical models for filtration through porous media interacting with the flow*, in Nonlinear Mathematical Problems in Industry I, M. Kawarada, N. Kenmochi, and N. Yanagihara, eds., Math. Sci. Appl. 1, Gakkotosho, Tokyo, 1993, pp. 61–85.

[10] A. Fasano and P. Tani, *Penetration of a wetting front in a porous medium with flux dependent hydraulic parameters*, in Nonlinear Problems in Applied Mathematics, T. S. Angell, L. P. Cook, R. E. Kleinman, and W. E. Olmstead, eds., SIAM, Philadelphia, PA, 1996, pp. 126–133.

[11] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.

[12] F. Talamucci, *Analysis of coupled heat-mass transport in freezing porous media*, Surveys Math. Indust., 7 (1997), pp. 93–139.

[13] F. Talamucci, *Flow through a porous medium with mass removal and diffusion*, NoDEA Nonlinear Differential Equations Appl., 5 (1998), pp. 427–444.

SIAM J. Math. Anal.
Vol. 31, No. 2, pp. 274–304

ANALYSIS AND CONSTRUCTION OF OPTIMAL MULTIVARIATE BIORTHOGONAL WAVELETS WITH COMPACT SUPPORT[*]

BIN HAN[†]

Abstract. In applications, it is well known that high smoothness, small support, and high vanishing moments are the three most important properties of a biorthogonal wavelet. In this paper, we shall investigate the mutual relations among these three properties. A characterization of $L_p\,(1 \le p \le \infty)$ smoothness of multivariate refinable functions is presented. It is well known that there is a close relation between a fundamental refinable function and a biorthogonal wavelet. We shall demonstrate that any fundamental refinable function, whose mask is supported on $[1-2r, 2r-1]^s$ for some positive integer r and satisfies the sum rules of optimal order $2r$, has L_p smoothness not exceeding that of the univariate fundamental refinable function with the mask b_r. Here the sequence b_r on $\mathbb{Z}$ is the unique univariate interpolatory refinement mask which is supported on $[1-2r, 2r-1]$ and satisfies the sum rules of order $2r$. Based on a similar idea, we shall prove that any orthogonal scaling function, whose mask is supported on $[0, 2r-1]^s$ for some positive integer r and satisfies the sum rules of optimal order r, has L_p smoothness not exceeding that of the univariate Daubechies orthogonal scaling function whose mask is supported on $[0, 2r-1]$. We also demonstrate that a similar result holds true for biorthogonal wavelets. Examples are provided to illustrate the general theory. Finally, a general CBC (cosets by cosets) algorithm is presented to construct all the dual refinement masks of any given interpolatory refinement mask with the dual masks satisfying arbitrary order of sum rules. Thus, for any scaling function which is fundamental, this algorithm can be employed to generate a dual scaling function with arbitrary approximation order. This CBC algorithm can be easily implemented. As a particular application of the general CBC algorithm, a TCBC (triangle cosets by cosets) algorithm is proposed. For any positive integer k and any interpolatory refinement mask a such that a is symmetric about all the coordinate axes, such a TCBC algorithm provides us with a dual mask of a such that the dual mask satisfies the sum rules of order $2k$ and is also symmetric about all the coordinate axes. As an application of this TCBC algorithm, a family of optimal bivariate biorthogonal wavelets is presented with the scaling function being a spline function.

Key words. biorthogonal wavelets, orthogonal wavelets, interpolatory subdivision schemes, fundamental functions, sum rules, L_p smoothness, critical exponent, algorithm

AMS subject classifications. 65D05, 41A25, 46E35, 41A05, 41A63, 41A30

PII. S0036141098336418

1. Introduction.

Based on the work [25, 26], the present paper deals with the analysis and construction of multivariate biorthogonal wavelets with some desired properties. It is well known that in various applications high smoothness, small support, and high vanishing moments are the three most important properties of a (bi)orthogonal wavelet. On the other hand, there is no C^∞ (bi)orthogonal wavelet with compact support. In this paper, we shall investigate the mutual relations among these three properties.

Compactly supported (bi)orthogonal wavelets on the real line have been found to be very useful in applications such as signal processing and image compression; for examples, see [1, 16, 36, 37]. In [10], Cohen, Daubechies, and Feauveau proposed a general way of constructing univariate biorthogonal wavelets. Although the tensor product (bi)orthogonal wavelets provide a family of multivariate (bi)orthogonal

[*] Received by the editors March 31, 1998; accepted for publication (in revised form) February 19, 1999; published electronically December 21, 1999. This work was supported by Killam Trust under the Isaak Walton Killam Memorial Scholarship.

http://www.siam.org/journals/sima/31-2/33641.html

[†] Program in Applied and Computational Mathematics, Department of Mathematics, Princeton University, Princeton, NJ 08544-1000 (bhan@math.princeton.edu, http://www.math.princeton.edu/~bhan).

wavelets to deal with problems in high dimensions in applications, it has its own advantages and disadvantages. Therefore, as noted in many papers [5, 9, 12, 24, 27, 37, 42] and references cited there, it is of interest in its own right to construct nontensor product (bi)orthogonal wavelets in the high dimensions. In the current literature, there are many papers on constructing multivariate biorthogonal wavelets, especially bivariate biorthogonal wavelets. To mention only a few here, see [9, 12, 24, 27, 37, 42] and references therein. Bivariate compactly supported quincunx biorthogonal wavelets were constructed by Cohen and Daubechies in [9]. In [42], a family of bivariate biorthogonal wavelets with the scaling function being a box spline was given by Riemenschneider and Shen.

Usually, a biorthogonal wavelet is derived from a multiresolution analysis generated by a pair consisting of a scaling function and its dual scaling function. The construction of wavelets in the multivariate setting is more challenging than its univariate counterpart; see [3, 10, 16, 24, 27, 33, 35, 38, 42] and references therein on construction of (bi)orthogonal wavelets from a multiresolution analysis. To obtain a biorthogonal wavelet, we have to find two refinable functions with some desired properties. A function ϕ is said to be *refinable* if it satisfies the following refinement equation:

$$\phi = \sum_{\beta\in\mathbb{Z}^s} a(\beta)\phi(2\cdot-\beta), \tag{1.1}$$

where a is a finitely supported sequence on $\mathbb{Z}^s$, called the *refinement mask*. If a satisfies $\sum_{\beta\in\mathbb{Z}^s} a(\beta) = 2^s$, then it is known (see [4]) that there exists a unique compactly supported distribution ϕ satisfying the refinement equation (1.1) subject to the condition $\widehat{\phi}(0) = 1$. This distribution is said to be the *normalized solution* of the refinement equation (1.1). Throughout this paper we shall use ϕ_a to denote the normalized solution of the refinement equation (1.1) with the mask a.

The concepts of linear independence and approximation order of a function play an important role in the study of biorthogonal wavelets. The shifts of a compactly supported function $\phi : \mathbb{R}^s \to \mathbb{C}$ are said to be *linearly independent* if for any $z \in \mathbb{C}^s$, there exists a multi-integer β in $\mathbb{Z}^s$ such that $\widehat{\phi}(z+2\pi\beta) \neq 0$. If for any $\xi \in \mathbb{R}^s$, there exists a multi-integer β in $\mathbb{Z}^s$ such that $\widehat{\phi}(\xi+2\pi\beta) \neq 0$, then the shifts of ϕ are said to be *stable*. See [34] for discussion on linear independence and stability.

By $\ell(\mathbb{Z}^s)$ we denote the linear space of all sequences on $\mathbb{Z}^s$. For a compactly supported function ϕ in $L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$, we define

$$S(\phi) := \left\{ \sum_{\alpha\in\mathbb{Z}^s} \phi(\cdot-\alpha)b(\alpha) \;:\; b \in \ell(\mathbb{Z}^s) \right\}$$

and call it the shift-invariant space generated by ϕ. For $h > 0$, the scaled space S^h is defined by $S^h := \{f(\cdot/h) \;:\; f \in S(\phi)\}$. For a positive integer k, we say that $S(\phi)$ provides *approximation order k* if for each sufficiently smooth function f in $L_p(\mathbb{R}^s)$, there exists a positive constant C such that

$$\inf_{g\in S^h} \|f-g\|_p \le Ch^k \qquad \forall\, h > 0.$$

The general procedure of constructing a biorthogonal wavelet is the following. First, find a refinable function ϕ in $L_2(\mathbb{R}^s)$ such that ϕ satisfies the refinement equation (1.1) with a finitely supported refinement mask a and the shifts of ϕ are linearly independent; such a function ϕ is called a *scaling function.* The next step is to find a refinable function ϕ^d in $L_2(\mathbb{R}^s)$ such that ϕ^d satisfies

$$\phi^d = \sum_{\beta\in\mathbb{Z}^s} a^d(\beta)\phi^d(2\cdot-\beta), \tag{1.2}$$

where a^d is a finitely supported sequence on $\mathbb{Z}^s$ and ϕ^d satisfies the following biorthogonal relation

$$\int_{\mathbb{R}^s} \overline{\phi(t-\alpha)}\phi^d(t)\,dt = \delta(\alpha) \qquad \forall\,\alpha\in\mathbb{Z}^s, \tag{1.3}$$

where $\delta(0)=1$ and $\delta(\alpha)=0\ \forall\,\alpha\in\mathbb{Z}^s\backslash\{0\}$. This function ϕ^d is called a *dual scaling function* of ϕ. If ϕ is the dual scaling function of itself, ϕ is called an *orthogonal scaling function.* Finally, a biorthogonal wavelet is derived from the above ϕ, ϕ^d, a, and a^d. The reader is referred to [5, 6, 7, 10, 16, 24, 27, 33, 35, 38, 42] for detail on the construction of a biorthogonal wavelet from a pair consisting of a scaling function and its dual scaling function. It is well known that the smoothness of the scaling function and its dual scaling function will determine the smoothness of their derived wavelets, and the approximation orders of the scaling function and its dual scaling function will determine the vanishing moments of their derived wavelets. For more detail on (bi)orthogonal wavelets, the reader is referred to [3, 5, 6, 7, 9, 10, 12, 13, 14, 16, 24, 27, 33, 35, 37, 42, 44] and references therein.

By Ω we denote the set of the vertices of the unit cube $[0,1]^s$. For a positive integer k, we say that a sequence a on $\mathbb{Z}^s$ satisfies the *sum rules* of order k if

$$\sum_{\beta\in\mathbb{Z}^s} a(2\beta+\varepsilon)p(2\beta+\varepsilon) = \sum_{\beta\in\mathbb{Z}^s} a(2\beta)p(2\beta) \qquad \forall\,\varepsilon\in\Omega,\ p\in\Pi_{k-1}, \tag{1.4}$$

where Π_{k-1} is the set of polynomials with total degree less than k. Let a function ϕ be a refinable function with a mask a. It was proved by Jia in [30, 31] that if the shifts of ϕ are stable, then $S(\phi)$ provides approximation order k if and only if the mask a satisfies the sum rules of order k. Therefore, it is evident that $S(\phi)$ (or $S(\phi^d)$) provides approximation order k if and only if the mask a (or a^d) satisfies the sum rules of order k.

Now it is natural to ask the following question: Given a scaling function with compact support, does a dual scaling function with compact support exist? As noted by Lemarié–Rieusset [39], the answer is yes at least in the univariate case. More precisely, given a scaling function with compact support, a dual scaling function always exists with compact support and arbitrarily high smoothness. Therefore, it is valuable to ask that for a scaling function, if we fix the size of the support of a dual scaling function, then what is the highest approximation order and the highest smoothness of a dual scaling function that we can expect? Based on our previous work on interpolatory subdivision schemes [25, 26], we shall answer the above question in this paper.

Here is an outline of this paper. In section 2, given a scaling function, we shall study the relation between the approximation order of its dual scaling function and the support of its dual scaling function. In section 3, a characterization of L_p smoothness of a multivariate refinable function is given. In section 4, we shall prove that any orthogonal scaling function, whose mask is supported on $[0,2r-1]^s$ ($r\in\mathbb{N}$) and satisfies the sum rules of optimal order r, has L_p smoothness not exceeding that of the univariate Daubechies orthogonal scaling function whose mask is supported on $[0, 2r-1]$. An example will be provided to illustrate our result. In section 5, we first study the optimal L_p smoothness of a fundamental refinable function if its mask is supported on $[1-2r, 2r-1]^s$ and satisfies the sum rules of optimal order $2r$. Next, for any given scaling function, we shall study the optimal smoothness of a dual scaling function if its support is fixed and it attains the optimal approximation order. Finally, in section 6, a general CBC (cosets by cosets) algorithm is presented to generate all the dual masks of a given interpolatory refinement mask. This algorithm can be easily implemented. In particular, as an application of this general construction, we shall propose a TCBC (triangle cosets by cosets) algorithm such that for any bivariate interpolatory mask which is symmetric about the two coordinate axes, we can construct a family of dual masks with arbitrary order of sum rules and symmetry about the two coordinate axes. At the end of this paper, a family of optimal bivariate biorthogonal wavelets is constructed from a spline scaling function.

2. Optimal approximation order of a dual scaling function. In this section, we shall first introduce some notation. For a given scaling function, we shall study the relation between the approximation order of a dual scaling function and the support of a dual scaling function.

In order to solve the refinement equation (1.1), we start with an initial function ϕ_0 given by

$$\phi_0(x_1, \dots, x_s) := \prod_{j=1}^{s} \chi(x_j), \qquad (x_1, \dots, x_s) \in \mathbb{R}^s,$$

where χ is the univariate hat function defined by

$$\chi(x) := \max\{1 - |x|, 0\}, \qquad x \in \mathbb{R}.$$

Then we employ the iteration scheme $Q_a^n \phi_0$, $n = 0, 1, 2, \dots$, where Q_a is the bounded linear operator on $L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$ given by

$$Q_a f := \sum_{\beta \in \mathbb{Z}^s} a(\beta) f(2 \cdot - \beta), \qquad f \in L_p(\mathbb{R}^s). \tag{2.1}$$

This iteration scheme is called a *subdivision scheme* or a *cascade algorithm* associated with the mask a (see [4, 17]). For any p such that $1 \le p \le \infty$, we say that the subdivision scheme associated with a mask a converges in the L_p norm if there exists a function f in $L_p(\mathbb{R}^s)$ such that $\lim_{n\to\infty} \|Q_a^n \phi_0 - f\|_p = 0$. If this is the case, then the limit function f must be the normalized solution of the refinement equation (1.1) with the refinement mask a.

Before proceeding further, we introduce some notation. By $\ell(\mathbb{Z}^s)$ we denote the space of all sequences on $\mathbb{Z}^s$ and by $\ell_0(\mathbb{Z}^s)$ the linear space of all finitely supported sequences on $\mathbb{Z}^s$. By δ we denote the element given by $\delta(0) = 1$ and $\delta(\beta) = 0$ $\forall\, \beta \in \mathbb{Z}^s \backslash \{0\}$. For $j = 1, \dots, s$, let e_j be the jth coordinate unit vector. The difference operator ∇_j on $\ell(\mathbb{Z}^s)$ is defined by $\nabla_j \lambda := \lambda - \lambda(\cdot - e_j)$, $\lambda \in \ell(\mathbb{Z}^s)$.

The *subdivision operator* associated with a mask a is defined by

$$S_a \lambda(\alpha) := \sum_{\beta \in \mathbb{Z}^s} a(\alpha - 2\beta) \lambda(\beta), \qquad \alpha \in \mathbb{Z}^s, \tag{2.2}$$

where $\lambda \in \ell_0(\mathbb{Z}^s)$. It was proved in [25] that the subdivision scheme associated with a mask a converges in the L_p norm if and only if

$$\lim_{n\to\infty} \max\big\{ \|\nabla_j S_a^n \delta\|_p^{1/n} \ : \ j = 1, \ldots, s \big\} < 2^{s/p}.$$

It is well known that there is a close relation between biorthogonal wavelets and fundamental refinable functions. A function ϕ is said to be *fundamental* if ϕ is continuous, $\phi(0) = 1$, and $\phi(\alpha) = 0 \ \forall\, \alpha \in \mathbb{Z}^s\backslash\{0\}$. If ϕ is a fundamental refinable function with a mask a, then it is necessary that

$$a(0) = 1 \quad \text{and} \quad a(2\beta) = 0 \qquad \forall\, \beta \in \mathbb{Z}^s\backslash\{0\}.$$

A mask that satisfies the above condition is called an *interpolatory refinement mask*.

The following fact is well known (see [8, 16, 40]) and reveals the relation between a biorthogonal wavelet and a fundamental refinable function.

LEMMA 2.1. *Let a function ϕ be a scaling function with a mask a and let ϕ^d be a dual scaling function of ϕ with a mask a^d. Define*

$$\Phi(x) := \int_{\mathbb{R}^s} \overline{\phi(t-x)}\phi^d(t)\, dt, \qquad x \in \mathbb{R}^s, \tag{2.3}$$

and

$$b(\alpha) := 2^{-s} \sum_{\beta\in\mathbb{Z}^s} \overline{a(\beta-\alpha)} a^d(\beta), \qquad \alpha \in \mathbb{Z}^s. \tag{2.4}$$

Then the function Φ is a fundamental refinable function satisfying the refinement equation (1.1) *with the interpolatory mask b. In other words, the mask a and a^d satisfy the following well-known discrete biorthogonal relation:*

$$\sum_{\beta\in\mathbb{Z}^s} \overline{a(\beta-2\alpha)}\, a^d(\beta) = 2^s\delta(\alpha) \qquad \forall\, \alpha \in \mathbb{Z}^s. \tag{2.5}$$

Conversely, if the masks a and a^d satisfy the above discrete biorthogonal relation (2.5) *and the subdivision schemes associated with a and a^d converge in the L_2 norm, respectively, then the functions ϕ and ϕ^d lie in $L_2(\mathbb{R}^s)$ and satisfy the biorthogonal relation* (1.3) *where the functions ϕ and ϕ^d are the normalized solutions of the refinement equations* (1.1) *with the masks a and a^d, respectively. Therefore, the function ϕ is a scaling function and ϕ^d is a dual scaling function of ϕ.*

If two sequences a and a^d on $\mathbb{Z}^s$ satisfy the discrete biorthogonal relation (2.5), then the mask a^d is called a *dual mask* of the mask a. Throughout this paper, we shall use the following notation:

$$\mathbb{T}^s := \{\, (z_1, \ldots, z_s) \in \mathbb{C}^s \ : \ |z_1| = \cdots = |z_s| = 1 \,\}.$$

For any sequence λ in $\ell_0(\mathbb{Z}^s)$, its *symbol* $\widetilde{\lambda}$ is given by

$$\widetilde{\lambda}(z) := \sum_{\beta\in\mathbb{Z}^s} \lambda(\beta) z^\beta, \qquad z \in \mathbb{T}^s. \tag{2.6}$$

By Lemma 2.1, we have $\widehat{\Phi}(\xi) = \overline{\widehat{\phi}(\xi)}\widehat{\phi^d}(\xi), \xi \in \mathbb{R}^s$ and $\widetilde{b}(z) = 2^{-s}\overline{\widetilde{a}(z)}\widetilde{a^d}(z), z \in \mathbb{T}^s$. The following result was proved in [26] and will be needed later.

THEOREM 2.2 (see [26, Theorems 2.1 and 2.2]). *Suppose that a is an interpolatory mask supported on $\mathbb{Z}^s \cap \Pi_{j=1}^s[-L_j, H_j]$ for some nonnegative integers L_j and H_j. If the mask a satisfies the sum rules of order k, then*

$$k \le \min_{1\le j\le s} \left(\left\lfloor \frac{L_j+1}{2} \right\rfloor + \left\lfloor \frac{H_j+1}{2} \right\rfloor \right),$$

where $\lfloor\cdot\rfloor$ is the floor function. Moreover, when $s = 1$, there exists a unique interpolatory refinement mask supported on $[-L_1, H_1]$ and satisfying the sum rules of order $\lfloor \frac{L_1+1}{2} \rfloor + \lfloor \frac{H_1+1}{2} \rfloor$.

By the above theorem, in the univariate case ($s = 1$), there is a unique interpolatory mask supported on $[1-2r, 2r-1]$ and satisfying the sum rules of order $2r$. This is the same interpolatory mask as given by Deslauriers and Dubuc in [18], and will be denoted by b_r throughout this paper. In the multivariate case ($s > 1$), such interpolatory masks are not unique. Let t_r be the sequence on $\mathbb{Z}^s$ given by

$$t_r(\alpha_1, \dots, \alpha_s) := b_r(\alpha_1) \cdots b_r(\alpha_s), \qquad (\alpha_1, \dots, \alpha_s) \in \mathbb{Z}^s. \tag{2.7}$$

Then t_r is a tensor product interpolatory refinement mask supported on $[1-2r, 2r-1]^s$ and it satisfies the sum rules of the optimal order $2r$.

Based on the above results, we have the following theorem.

THEOREM 2.3. *Let ϕ be a scaling function with its refinement mask a supported on $\Pi_{j=1}^s[-l_j, h_j]$ for some nonnegative integers l_j and h_j, and let ϕ^d be its dual scaling function with its mask a^d supported on $\Pi_{j=1}^s[-L_j, H_j]$ for some nonnegative integers L_j and H_j. Suppose that a satisfies the sum rules of order k; then a^d can satisfy the sum rules of order at most*

$$\min_{1\le j\le s} \left(\left\lfloor \frac{h_j+L_j+1}{2} \right\rfloor + \left\lfloor \frac{l_j+H_j+1}{2} \right\rfloor \right) - k,$$

where $\lfloor\cdot\rfloor$ is the floor function.

Proof. Let b be the sequence defined in (2.4). Then by Lemma 2.1, b is an interpolatory mask and b is supported on $\Pi_{j=1}^s[-h_j - L_j, l_j + H_j]$. From Theorem 2.2, we see that b can satisfy the sum rules of order at most $\min_{1\le j\le s} \left(\lfloor \frac{h_j+L_j+1}{2} \rfloor + \lfloor \frac{l_j+H_j+1}{2} \rfloor \right)$. To complete the proof, it suffices to prove that if the mask a^d satisfies the sum rules of order $\widetilde{k}$, then b will satisfy the sum rules of order at least $k + \widetilde{k}$. Denote

$$\mathbb{Z}_+^s := \{ (\alpha_1, \dots, \alpha_s) \in \mathbb{Z}^s \ : \ \alpha_j \ge 0 \quad \forall\, j = 1, \dots, s \},$$

and $|\mu| := \mu_1 + \cdots + \mu_s$ for $\mu = (\mu_1, \dots, \mu_s) \in \mathbb{Z}_+^s$ and $\alpha^\mu := \alpha_1^{\mu_1} \cdots \alpha_s^{\mu_s}$ for $\alpha = (\alpha_1, \dots, \alpha_s) \in \mathbb{Z}^s$. Thus, by the definition of the sum rules given in (1.4), it suffices to prove that

$$\sum_{\alpha\in\mathbb{Z}^s} b(2\alpha+\varepsilon)(2\alpha+\varepsilon)^\mu = \sum_{\alpha\in\mathbb{Z}^s} b(2\alpha)(2\alpha)^\mu \qquad \forall\, \mu \in \mathbb{Z}_+^s,\, |\mu| < k + \widetilde{k},\, \varepsilon \in \Omega.$$

By the definition of the sequence b given in (2.4), we can rewrite the above equality as follows: for any $\varepsilon \in \Omega$ and $\mu \in \mathbb{Z}_+^s$ such that $|\mu| < k + \widetilde{k}$,

$$\sum_{\alpha\in\mathbb{Z}^s} \sum_{\beta\in\mathbb{Z}^s} \overline{a(\beta - 2\alpha - \varepsilon)} a^d(\beta)(2\alpha+\varepsilon)^\mu = \sum_{\alpha\in\mathbb{Z}^s} \sum_{\beta\in\mathbb{Z}^s} \overline{a(\beta - 2\alpha)} a^d(\beta)(2\alpha)^\mu.$$

Therefore, it suffices to prove that the left side of the above equality

$$C_\varepsilon := \sum_{\varepsilon' \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \sum_{\beta \in \mathbb{Z}^s} \overline{a(2\beta + \varepsilon' - 2\alpha - \varepsilon)} a^d(2\beta + \varepsilon')(2\alpha + \varepsilon)^\mu$$

does not depend on ε for any $\mu \in \mathbb{Z}_+^s$ such that $|\mu| < k + \widetilde{k}$. On the other hand,

$$\begin{aligned}(2\alpha + \varepsilon)^\mu &= \big((2\beta + \varepsilon') - (2\beta + \varepsilon' - 2\alpha - \varepsilon)\big)^\mu \\ &= \sum_{0 \le \nu \le \mu} (-1)^{|\nu|} \frac{\mu!}{\nu!(\mu - \nu)!} (2\beta + \varepsilon' - 2\alpha - \varepsilon)^\nu (2\beta + \varepsilon')^{\mu - \nu},\end{aligned}$$

where $\mu! := \mu_1! \cdots \mu_s!$ for $\mu = (\mu_1, \ldots, \mu_s)$ and $\nu \le \mu$ if and only if $\nu_j \le \mu_j \ \forall\, j = 1, \ldots, s$. Thus, C_ε can be rewritten as

$$\begin{aligned}C_\varepsilon = \sum_{0 \le \nu \le \mu} (-1)^{|\nu|} \frac{\mu!}{\nu!(\mu - \nu)!} &\sum_{\beta \in \mathbb{Z}^s} \sum_{\varepsilon' \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \\ &\overline{a(2\beta + \varepsilon' - 2\alpha - \varepsilon)} (2\beta + \varepsilon' - 2\alpha - \varepsilon)^\nu a^d(2\beta + \varepsilon')(2\beta + \varepsilon')^{\mu - \nu}.\end{aligned}$$

Therefore, it suffices to demonstrate that for any $\nu \in \mathbb{Z}_+^s$ such that $0 \le \nu \le \mu$,

$$C_{\varepsilon,\nu} := \sum_{\varepsilon' \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \overline{a(2\alpha + \varepsilon' - \varepsilon)} (2\alpha + \varepsilon' - \varepsilon)^\nu \sum_{\beta \in \mathbb{Z}^s} a^d(2\beta + \varepsilon')(2\beta + \varepsilon')^{\mu - \nu}$$

does not depend on ε. Note that $|\mu - \nu| + |\nu| = |\mu| < k + \widetilde{k}$ implies that either $|\mu - \nu| < k$ or $|\nu| < \widetilde{k}$. If $|\nu| < k$, then $\sum_{\alpha \in \mathbb{Z}^s} a(2\alpha + \varepsilon' - \varepsilon)(2\alpha + \varepsilon' - \varepsilon)^\nu$ does not depend on both ε and ε' since the sequence a satisfies the sum rules of order k. Hence, for any ν in $\mathbb{Z}_+^s$ such that $|\nu| < k$, we have

$$C_{\varepsilon,\nu} = \overline{\sum_{\alpha \in \mathbb{Z}^s} a(2\alpha)(2\alpha)^\nu} \sum_{\beta \in \mathbb{Z}^s} a^d(\beta)\beta^{\mu - \nu}$$

does not depend on ε. Similarly, if $|\mu - \nu| < \widetilde{k}$, then $\sum_{\beta \in \mathbb{Z}^s} a^d(2\beta + \varepsilon')(2\beta + \varepsilon')^{\mu - \nu}$ does not depend on ε' since the sequence a^d satisfies the sum rules of order $\widetilde{k}$. Therefore,

$$C_{\varepsilon,\nu} = \sum_{\beta \in \mathbb{Z}^s} a^d(2\beta)(2\beta)^{\mu - \nu} \overline{\sum_{\alpha \in \mathbb{Z}^s} a(\alpha)\alpha^\nu}$$

does not depend on ε which completes the proof. □

From the proof of Theorem 2.3, it is straightforward to obtain the following result.

Corollary 2.4. *If a function ϕ in $L_2(\mathbb{R}^s)$ is an orthogonal scaling function with its mask a supported on $[0, r]^s$ for some positive integer r, then the mask a can satisfy the sum rules of order at most $\lfloor \frac{r+1}{2} \rfloor$. Therefore, $S(\phi)$ can provide approximation order at most $\lfloor \frac{r+1}{2} \rfloor$.*

3. Characterization of L_p smoothness of a refinable function. In this section, we will study the smoothness of a refinable function in the multivariate setting. Many results on the analysis of L_2 smoothness of a refinable function both in the univariate case and in the multivariate case are obtained in the current literature. To mention only a few here, see [11, 17, 23, 29, 41, 43, 45] and references therein. For $s = 1$, the characterization of L_p smoothness was given by Villemoes in [45].

In this section, based on a result of Ditzian [19, 20], we present a simple proof to characterize the L_p smoothness of a multivariate refinable function. Jia will discuss the L_p smoothness of a refinable function with an arbitrary dilation matrix in a forthcoming paper [32].

We shall use the generalized Lipschitz space to measure smoothness of a given function. For any vector y in $\mathbb{R}^s$, the *difference operator* ∇_y on $L_p(\mathbb{R}^s)$ is defined to be

$$\nabla_y f = f - f(\cdot - y), \qquad f \in L_p(\mathbb{R}^s).$$

Let k be a positive integer. The kth *modulus of smoothness* of a function f in $L_p(\mathbb{R}^s)$ is defined by

$$\omega_k(f,h)_p := \sup_{|y|\le h} \|\nabla_y^k f\|_p, \qquad h > 0.$$

For $\nu > 0$, let k be an integer greater than ν. The *generalized Lipschitz space* $Lip^*\big(\nu, L_p(\mathbb{R}^s)\big)$ consists of those functions f in $L_p(\mathbb{R}^s)$ for which

$$\omega_k(f,h)_p \le Ch^\nu \qquad \forall\, h > 0, \tag{3.1}$$

where C is a constant independent of h, or in other words, $\omega_k(f,h)_p = O(h^\nu)$.

The L_p smoothness of a function $f \in L_p(\mathbb{R}^s)$ in the L_p norm sense is described by its L_p *critical exponent* $\nu_p(f)$ defined by

$$\nu_p(f) := \sup\left\{ \nu \ : \ f \in Lip^*\big(\nu, L_p(\mathbb{R}^s)\big) \right\}. \tag{3.2}$$

In the following, we will characterize the L_p $(1 \le p \le \infty)$ smoothness of a refinable function in multidimensional spaces. To do this, we need the following result on moduli of smoothness, which is based on a result of Ditzian in [19, 20].

Theorem 3.1. *Let f be a function in $L_p(\mathbb{R}^s)$ and ν be a positive real number. Then f belongs to the space $Lip^*\big(\nu, L_p(\mathbb{R}^s)\big)$ if and only if for an integer k greater than ν, there exists a positive constant C such that*

$$\max\{\, \|\nabla_{2^{-n}e_i}^k f\|_p \ : \ i = 1,\ldots,s \,\} \le C2^{-n\nu} \qquad \forall n \in \mathbb{N}, \tag{3.3}$$

where e_i is the ith coordinate unit vector.

Proof. Necessity: If f belongs to $Lip^*\big(\nu, L_p(\mathbb{R}^s)\big)$, then by the definition of the Lipschitz space $Lip^*\big(\nu, L_p(\mathbb{R}^s)\big)$, there exists a positive constant C such that

$$\|\nabla_{2^{-n}e_i}^k f\|_p \le \omega_k(f, 2^{-n}) \le C2^{-n\nu} \qquad \forall\, 1 \le i \le s, n \in \mathbb{N}.$$

Hence inequality (3.3) holds true.

Sufficiency: If inequality (3.3) holds true, then we can demonstrate that there exists a positive constant C_1 such that

$$\|\nabla_{he_i}^k f\|_p \le C_1 h^\nu \qquad \forall\, 1 \le i \le s, h > 0. \tag{3.4}$$

Let g be a simple function such that $\|g\|_q = 1$, where $1/p + 1/q = 1$. Define

$$F(x) := f * g(x) = \int_{\mathbb{R}^s} f(x-t)g(t)\,dt, \qquad x \in \mathbb{R}^s.$$

Then the function F is continuous and bounded. Note that the inequality (3.3) implies that for any $i = 1, \ldots, s$,

$$\|\nabla^k_{2^{-n}e_i} F\|_\infty = \|(\nabla^k_{2^{-n}e_i} f) * g\|_\infty \le \|\nabla^k_{2^{-n}e_i} f\|_p \|g\|_q \le C 2^{-n\nu} \qquad \forall\, n \in \mathbb{N}.$$

Therefore, in particular, we have

$$|\nabla^k_{2^{-n}e_i} F(te_i)| \le C 2^{-n\nu} \qquad \forall\, t \in \mathbb{R}, n \in \mathbb{N}.$$

By a result of Boman [2, Theorem 1] and Ditzian [20], there exists a positive constant C_1 depending only on k and C (independent of g) such that

$$|\nabla^k_{he_i} F(te_i)| \le C_1 h^\nu \qquad \forall\, t \in \mathbb{R},\, h > 0. \tag{3.5}$$

Note that $\nabla^k_{he_i} F(0) = (\nabla^k_{he_i} f) * g(0)$. It follows from the above inequality (3.5) that for any simple function g with $\|g\|_q = 1$, we have that for any $i = 1, \ldots, s$,

$$\left| \int_{\mathbb{R}^s} \left(\nabla^k_{he_i} f\right)(-x) g(x)\, dx \right| = |(\nabla^k_{he_i} f) * g(0)| = |\nabla^k_{he_i} F(0)| \le C_1 h^\nu \quad \forall\, h > 0.$$

This yields

$$\|\nabla^k_{he_i} f\|_p = \sup_{\|g\|_q = 1} \left| \int_{\mathbb{R}^s} \left(\nabla^k_{he_i} f\right)(-x) g(x)\, dx \right| \le C_1 h^\nu \qquad \forall\, 1 \le i \le s, h > 0.$$

Therefore, inequality (3.4) is verified. By inequality (3.4) and a result of Ditzian [19, Corollary 5.2, and also cf. Theorem 5.1], it is straightforward to see that the function f belongs to the function space $Lip^*\big(\nu, L_p(\mathbb{R}^s)\big)$. ◻

Remark 3.2. In fact, the result in Corollary 5.2 of Ditzian [19] is a Marchaud-type inequality which says that to characterize the kth modulus of smoothness of a function in $L_p(\mathbb{R}^s)$ in the L_p norm sense, the information of the kth modulus of smoothness in s independent directions is enough. More precisely, for any vector y in $\mathbb{R}^s$, we denote $\omega_k(f, h, y)_p := \sup_{|t| \le h} \|\nabla^k_{ty} f\|_p$, $h > 0$. Let $y_i, i = 1, \ldots, s$ be s linearly independent vectors in $\mathbb{R}^s$. Then for any $\nu > 0$ and an integer $k > \nu$, $\omega_k(f, h)_p = O(h^\nu)$ if and only if $\omega_k(f, h, y_i)_p = O(h^\nu)$ $\forall\, i = 1, \ldots, s$. Therefore, in Theorem 3.1, the vectors $e_i, i = 1, \ldots, s$ can be replaced by vectors $y_i, i = 1, \ldots, s$ provided that $y_i, i = 1, \ldots, s$ are linearly independent vectors in $\mathbb{R}^s$. For more detail on the above result, the reader is referred to the works of Boman [2] and Ditzian [19, 20].

Based on the above result, the following theorem gives us a characterization of the critical exponent $\nu_p(\phi)$ of a refinable function ϕ in $L_p(\mathbb{R}^s)$ in terms of its mask provided that the shifts of the refinable function ϕ are stable.

THEOREM 3.3. *Let a function ϕ in $L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$ be the normalized solution of the refinement equation* (1.1) *with a finitely supported refinement mask a on $\mathbb{Z}^s$ such that $\sum_{\beta \in \mathbb{Z}^s} a(\beta) = 2^s$. For any nonnegative integer k, let*

$$\sigma_{k,p}(a) := \lim_{n \to \infty} \max\{\, \|\nabla^k_i S^n_a \delta\|_p^{1/n} \;:\; i = 1, \ldots, s \,\}.$$

Then

$$\min\{\, k,\ \nu_p(\phi)\,\} \ge s/p - \log_2 \sigma_{k,p}(a). \tag{3.6}$$

In addition, if the shifts of ϕ are stable, then

$$\min\{\, k, \nu_p(\phi)\,\} = s/p - \log_2 \sigma_{k,p}(a). \tag{3.7}$$

More generally, let $Y := \{\, y_i \in \mathbb{Z}^s \;:\; i = 1, \ldots, s \,\}$ be a set of s linearly independent vectors. Define

$$\sigma_{k,p,Y}(a) := \lim_{n\to\infty} \max\{\, \|\nabla^k_{y_i} S^n_a \delta\|_p^{1/n} \;:\; i = 1, \ldots, s \,\}.$$

Then the above results still hold true if $\sigma_{k,p}(a)$ is replaced with $\sigma_{k,p,Y}(a)$.

Proof. By the definition of $\sigma_{k,p}(a)$, for any real number r such that $r > \sigma_{k,p}(a)$, there exists a positive constant C_r such that

$$\max\{\, \|\nabla^k_i\, S^n_a \delta\|_p \;:\; i = 1, \ldots, s \,\} \le C_r r^n \qquad \forall\, n \in \mathbb{N}. \tag{3.8}$$

By induction and the definition of the subdivision operator defined in (2.2), we observe that

$$\nabla^k_{2^{-n}e_i} \phi = \sum_{\beta\in\mathbb{Z}^s} \nabla^k_i S^n_a \delta(\beta)\phi(2^n \cdot -\beta), \qquad i = 1, \ldots, s. \tag{3.9}$$

Since the function ϕ in $L_p(\mathbb{R}^s)$ is compactly supported, from (3.9), there exists a positive constant C_1 depending only on ϕ such that

$$\|\nabla^k_{2^{-n}e_i}\, \phi\|_p \le C_1 2^{-ns/p} \|\nabla^k_i\, S^n_a \delta\|_p \qquad \forall\, n \in \mathbb{N}, i = 1, \ldots, s.$$

Therefore, it follows from inequality (3.8) that

$$\|\nabla^k_{2^{-n}e_i}\, \phi\|_p \le C_1 C_r 2^{-ns/p} r^n \qquad \forall\, n \in \mathbb{N}, i = 1, \ldots, s. \tag{3.10}$$

On the other hand, by induction, we observe $\sigma_{k,p}(a) \ge 2^{s/p-k}$ since $\sum_{\beta\in\mathbb{Z}^s} a(\beta) = 2^s$. Therefore, the inequality $k \ge s/p - \log_2 \sigma_{k,p}(a)$ holds true for any nonnegative integer k. Since $r > \sigma_{k,p}(a)$, we deduce that $k \ge s/p - \log_2 \sigma_{k,p}(a) > s/p - \log_2 r$. By Theorem 3.1, it follows from inequality (3.10) that $\phi \in Lip^*\big(s/p - \log_2 r, L_p(\mathbb{R}^s)\big)$ for any r such that $r > \sigma_{k,p}(a)$. Thus in conclusion, we have

$$\min\{\, k, \nu_p(\phi) \,\} \ge s/p - \log_2 \sigma_{k,p}(a).$$

If the shifts of the function ϕ are stable, to prove (3.7), it suffices to prove that $\min\{k, \nu_p(\phi)\} \le s/p - \log_2 \sigma_{k,p}(a)$, equivalently, it suffices to prove that

$$\sigma_{k,p}(a) \le 2^{s/p-\nu} \qquad \forall\, 0 < \nu < \min\{\, k,\, \nu_p(\phi) \,\}.$$

Since the shifts of the function ϕ are stable and ϕ lies in $L_p(\mathbb{R}^s)$, from (3.9), there exists a positive constant C_2 depending only on the function ϕ such that

$$\|\nabla^k_i\, S^n_a \delta\|_p \le C_2 2^{ns/p} \|\nabla^k_{2^{-n}e_i}\, \phi\|_p \qquad \forall\, n \in \mathbb{N},\, i = 1, \ldots, s.$$

Since $\phi \in Lip^*\big(\nu, L_p(\mathbb{R}^s)\big)$ and $k > \nu$, by Theorem 3.1, we have

$$\max_{1\le i\le s}\{\, \|\nabla^k_i\, S^n_a \delta\|_p \,\} \le C_2 2^{ns/p} \max_{1\le i\le s}\{\, \|\nabla^k_{2^{-n}e_i}\, \phi\|_p \,\} \le C_2 C 2^{n(s/p-\nu)} \qquad \forall\, n \in \mathbb{N}.$$

Therefore, the inequality $\sigma_{k,p}(a) \le 2^{s/p-\nu}$ holds true, as desired. The last assertion of this theorem follows directly from Remark 3.2. □

Remark 3.4. If the shifts of the function ϕ are stable and its mask a satisfies the sum rules of order k but not $k + 1$, then $\nu_p(\phi) \le k$ (see [4, 30]) and therefore, by

Theorem 3.3, $\nu_p(\phi) = s/p - \log_2 \sigma_{k,p}(a)$. Another remark about the above theorem is that by carefully choosing the set Y, the equality in (3.6) may hold even when the shifts of the function are not stable. For example, let

$$\phi(x) = \max\{1 - |x|/2,\, 0\}, \qquad x \in \mathbb{R}.$$

Then the function ϕ is a refinable function with its mask a given by its symbol $\widetilde{a}(z) := 1 + (z^{-2} + z^2)/2$. It is a known fact that the shifts of ϕ are *not* stable and $\nu_p(\phi) = 1 + 1/p$ for any p such that $1 \le p \le \infty$. On the other hand, choose $y = 2$. It is not difficult to verify that $\sigma_{2,p,y}(a) := \lim_{n\to\infty} \|\nabla_y^2 S_a^n \delta\|_p^{1/n} = 1/2$. Therefore, we still have $\nu_p(\phi) = 1/p - \log_2 \sigma_{2,p,y}(a) = 1/p + 1$ for any p such that $1 \le p \le \infty$. In passing, we mention that $\sigma_{k,2}(a)$ can be obtained by finding the spectral radius of a finite matrix by [25, Theorem 4.1]. If $\sigma_{k,p}(a) < 2^{s/p}$ for some positive integer k, then $\sigma_{1,p}(a) < 2^{s/p}$ and therefore, by [25, Theorem 3.2] the subdivision scheme associated with the mask a converges in the L_p norm and $\phi_a \in L_p(\mathbb{R}^s)$.

Finally, in this section, we prove the following result which will be needed later.

THEOREM 3.5. *Suppose that a function ϕ is a fundamental real-valued function on the real line and ϕ satisfies the refinement equation* (1.1) *with an interpolatory refinement mask a supported on* $[-3, 3]$. *Then the inequality $\nu_\infty(\phi) \le 2$ holds true and therefore, $\phi \notin C^2(\mathbb{R})$.*

Proof. We use proof by contradiction to verify our claim. Suppose $\nu_\infty(\phi) > 2$. Then a must satisfy the sum rules of order at least 3 (see [4, 30]). By a simple calculation, it is not difficult to see that the symbol $\widetilde{a}(z)$ can be written as

$$\widetilde{a}(z) = z^{-3}\,(1+z)^3\,\widetilde{c}(z) \quad \text{with} \quad \widetilde{c}(z) := t - 3\,t\,z + (3/8 + 3t)\,z^2 - (1/8 + t)\,z^3,$$

for some $t \in \mathbb{R}$. By [28, Theorem 3.2] or [26, Theorem 3.1], we observe that

$$\sigma_{3,\infty}(a) = \sigma_{0,\infty}(c) = \lim_{n\to\infty} \big(\max\big\{\|B_1 \cdots B_n\| \;:\; B_1, \ldots, B_n \in \{A_0, A_1\}\big\}\big)^{1/n},$$

where A_0 and A_1 are matrices given by

$$A_0 := \begin{pmatrix} t & 3/8 + 3\,t & 0 \\ 0 & -3\,t & -1/8 - t \\ 0 & t & 3/8 + 3\,t \end{pmatrix}$$

and

$$A_1 := \begin{pmatrix} -3\,t & -1/8 - t & 0 \\ t & 3/8 + 3\,t & 0 \\ 0 & -3\,t & -1/8 - t \end{pmatrix}.$$

Therefore, it is evident that

$$\sigma_{3,\infty}(a) = \sigma_{0,\infty}(c) \ge \lim_{n\to\infty} \|A_0^n\|^{1/n} =: \rho(A_0),$$

where $\rho(A_0)$ is the spectral radius of A_0. By a simple calculation again, we see that $\lambda = 3/16 + \sqrt{(3/8)^2 + 4(t + 8t^2)}/2$ is an eigenvalue of A_0. Note that

$$\lambda = 3/16 + \sqrt{1/256 + 8(t + 1/16)^2} \ge 1/4 \qquad \forall\, t \in \mathbb{R}.$$

This yields

$$\sigma_{3,\infty}(a) = \sigma_{0,\infty}(c) \ge \rho(A_0) \ge 1/4.$$

Since the function ϕ is a fundamental function, the shifts of ϕ are stable. By Theorem 3.3, we have

$$\min\{\,3,\,\nu_\infty(\phi)\,\} = -\log_2 \sigma_{3,\infty}(a) \le -\log_2(1/4) = 2.$$

This is a contradiction to our assumption $\nu_\infty(\phi) > 2$. Hence, the inequality $\nu_\infty(\phi) \le 2$ holds true. This completes our proof. □

4. Optimal orthogonal wavelets in the multivariate setting. In [15], Daubechies first constructed a family of compactly supported orthogonal scaling functions on the real line, namely, ϕ_{D_r} $(r \in \mathbb{N})$, where ϕ_{D_r} satisfies the refinement equation (1.1) with the mask D_r supported on $[0, 2r-1]$. It is observed (see [40]) that D_r satisfies the sum rules of order r and $\overline{\widetilde{D_r}(z)}\widetilde{D_r}(z) = 2\widetilde{b}_r(z)$ for any z in $\mathbb{T}$, where b_r is the unique univariate interpolatory mask which is supported on $[1-2r, 2r-1]$ and satisfies the sum rules of order $2r$. Therefore, by Corollary 2.4, the mask D_r attains the sum rules of optimal order r. In the multivariate setting, due to the lack of the Riesz factorization theorem, it is much more difficult to construct multivariate orthogonal scaling functions than to construct univariate ones. In the current literature, there are few examples of nontensor product multivariate orthogonal scaling functions.

Before proceeding further, we need the following two lemmas.

LEMMA 4.1. *Let a sequence a on $\mathbb{Z}^s$ be an interpolatory mask supported on $[1-2r, 2r-1]^s$ for some positive integer r. Define a new sequence a_1 on $\mathbb{Z}$ as follows:*

$$a_1(k) = 2^{1-s} \sum_{\alpha_2 \in \mathbb{Z}} \cdots \sum_{\alpha_s \in \mathbb{Z}} a(k, \alpha_2, \ldots, \alpha_s), \qquad k \in \mathbb{Z}. \tag{4.1}$$

If the mask a satisfies the sum rules of order at least $2r-1$, then a_1 is a univariate interpolatory refinement mask satisfying the sum rules of order $2r-1$. Moreover, if the mask a satisfies the sum rules of order $2r$, then the mask a_1 must be the mask b_r, the unique interpolatory refinement mask which is supported on $[1-2r, 2r-1]$ and satisfies the sum rules of order $2r$.

Proof. By the definition of sum rules given in (1.4), it is easily seen that the sequence a_1 satisfies the same order of sum rules as the sequence a does. Hence, to complete the proof, it suffices to prove that a_1 is a univariate interpolatory refinement mask. Namely, we have to prove that $a_1(2k) = 0\ \forall\, k \in \mathbb{Z}\backslash\{0\}$. To this end, it suffices to prove that for any ε in Ω such that $\varepsilon = (0, \varepsilon_2, \ldots, \varepsilon_s)$,

$$\sum_{\alpha_2 \in \mathbb{Z}} \cdots \sum_{\alpha_s \in \mathbb{Z}} a(2k, 2\alpha_2 + \varepsilon_2, \ldots, 2\alpha_s + \varepsilon_s) = 0 \qquad \forall\, k \in \mathbb{Z}\backslash\{0\}. \tag{4.2}$$

Let b be a sequence on $\mathbb{Z}$ given by

$$b(k) := \sum_{\alpha_2 \in \mathbb{Z}} \cdots \sum_{\alpha_s \in \mathbb{Z}} a(2k, 2\alpha_2 + \varepsilon_2, \ldots, 2\alpha_s + \varepsilon_s), \qquad k \in \mathbb{Z}.$$

It is evident that b is supported on $[1-r, r-1]$ since a is supported on $[1-2r, 2r-1]^s$. Note that the mask a is an interpolatory refinement mask which satisfies the sum rules of order $2r-1$. By the definition of sum rules given in (1.4), for any integer j such that $0 \le j < 2r-1$, we deduce that

$$\sum_{k \in \mathbb{Z}} b(k)(2k)^j = \sum_{k \in \mathbb{Z}} \sum_{\alpha_2 \in \mathbb{Z}} \cdots \sum_{\alpha_s \in \mathbb{Z}} a(2k, 2\alpha_2 + \varepsilon_2, \ldots, 2\alpha_s + \varepsilon_s)(2k)^j = \delta(j).$$

This gives us

$$\sum_{k=1-r}^{r-1} b(k)k^j = \delta(j), \qquad 0 \le j < 2r-1. \tag{4.3}$$

This linear system has $2r-1$ unknowns $b(1-r), \dots, b(r-1)$ and $2r-1$ equations and its coefficient matrix is a Vandermonde matrix. Hence, it has a unique solution. It is easily seen that $b(j) = \delta(j)$, $j = 1-r, \dots, r-1$, is a solution to the above linear system. This verifies (4.2), thereby completing the proof. □

LEMMA 4.2. *Let a function ϕ in $L_p(\mathbb{R}^s)$ $(1 \le p \le \infty)$ be the normalized solution of the refinement equation* (1.1) *with a finitely supported refinement mask a on $\mathbb{Z}^s$. Let the sequence a_1 be given by* (4.1) *and ϕ_{a_1} be the normalized solution of the refinement equation* (1.1) *with the refinement mask a_1. Suppose that the shifts of ϕ are stable. Then the subdivision scheme associated with the mask a_1 converges in the L_p norm and $\nu_p(\phi) \le \nu_p(\phi_{a_1})$.*

Proof. In the following, we shall prove that for any nonnegative integer k, there exists a positive constant C such that

$$\|\nabla_1^k S_{a_1}^n \delta\|_p \le C 2^{n(1-s)/p} \|\nabla_1^k S_a^n \delta\|_p \qquad \forall\, n \in \mathbb{N}. \tag{4.4}$$

From the definition of the subdivision operator given in (2.2), we observe that $\widetilde{S_a^n\delta}(z) = \prod_{j=0}^{n-1} \widetilde{a}(z^{2^j})$ for any z in $\mathbb{T}^s$. Therefore, we deduce $\widetilde{S_{a_1}^n\delta}(z) = 2^{(1-s)n}\widetilde{S_a^n\delta}(z,1,\dots,1)$ for any z in $\mathbb{T}$ since $\widetilde{a_1}(z) = 2^{1-s}\widetilde{a}(z,1,\dots,1)$ and $\widetilde{S_{a_1}^n\delta}(z) = \prod_{j=0}^{n-1}\widetilde{a_1}(z^{2^j})$ for any z in $\mathbb{T}$. That is,

$$S_{a_1}^n\delta(j) = 2^{(1-s)n} \sum_{\beta_2\in\mathbb{Z}} \cdots \sum_{\beta_s\in\mathbb{Z}} S_a^n\delta(j,\beta_2,\dots,\beta_s) \qquad \forall\, j \in \mathbb{Z},\ n \in \mathbb{N}. \tag{4.5}$$

Since $\nabla_1\lambda(\beta) = \lambda(\beta) - \lambda(\beta - e_1)$, $\lambda \in \ell_0(\mathbb{Z}^s)$, where e_1 is the first coordinate unit vector, we have

$$\begin{aligned}\nabla_1^k S_{a_1}^n\delta(j) &= 2^{(1-s)n}\nabla_1^k \sum_{\beta_2\in\mathbb{Z}} \cdots \sum_{\beta_s\in\mathbb{Z}} S_a^n\delta(j,\beta_2,\dots,\beta_s)\\ &= 2^{(1-s)n} \sum_{\beta_2\in\mathbb{Z}} \cdots \sum_{\beta_s\in\mathbb{Z}} \nabla_1^k S_a^n\delta(j,\beta_2,\dots,\beta_s).\end{aligned}$$

Since the mask a is finitely supported, there exists a positive integer r such that $\operatorname{supp} a \subseteq [-r,r]^s$. It is easily seen that $\operatorname{supp} S_a^n\delta \subseteq [-2^n r, 2^n r]$. Therefore, the above equality can be rewritten as

$$\nabla_1^k S_{a_1}^n\delta(j) = 2^{(1-s)n} \sum_{\beta_2=-2^n r}^{2^n r} \cdots \sum_{\beta_s=-2^n r}^{2^n r} \nabla_1^k S_a^n\delta(j,\beta_2,\dots,\beta_s), \qquad j \in \mathbb{Z}.$$

Applying the Hölder inequality to the above sum, we obtain

$$\begin{aligned}|\nabla_1^k S_{a_1}^n\,\delta(j)|^p &\le 2^{n(1-s)p}(2^{n+1}r+1)^{(s-1)p/q} \sum_{\beta_2\in\mathbb{Z}} \cdots \sum_{\beta_s\in\mathbb{Z}} |\nabla_1^k\, S_a^n\,\delta(j,\beta_2,\dots,\beta_s)|^p\\ &\le C_1 2^{n(1-s)} \sum_{\beta_2\in\mathbb{Z}} \cdots \sum_{\beta_s\in\mathbb{Z}} |\nabla_1^k\, S_a^n\,\delta(j,\beta_2,\dots,\beta_s)|^p,\end{aligned}$$

where $1/p + 1/q = 1$ and $C_1 = (2r+1)^{(s-1)p/q}$. It follows from the above inequality that

$$\|\nabla_1^k S_{a_1}^n \delta\|_p \le C_1^{1/p} 2^{n(1-s)/p} \|\nabla_1^k S_a^n \delta\|_p \qquad \forall\, n \in \mathbb{N}.$$

Therefore, the inequality (4.4) holds true. Since the shifts of ϕ are stable and ϕ lies in $L_p(\mathbb{R}^s)$, the subdivision scheme associated with the mask a converges in the L_p norm. That is, by [25, Theorem 3.2], it is equivalent to

$$\lim_{n\to\infty} \max\{\, \|\nabla_i S_a^n \delta\|_p^{1/n} \;:\; i = 1, \dots, s \,\} < 2^{s/p}.$$

The reader is referred to [25] for a detailed discussion on the convergence of a subdivision scheme in the L_p norm. Taking $k = 1$ in (4.4), we get

$$\lim_{n\to\infty} \|\nabla_1 S_{a_1}^n \delta\|_p^{1/n} \le 2^{(1-s)/p} \lim_{n\to\infty} \max\{\, \|\nabla_i S_a^n \delta\|_p^{1/n} \;:\; i = 1, \dots, s \,\} < 2^{1/p}.$$

Hence, the subdivision scheme associated with the mask a_1 converges in the L_p norm. In particular, we have $\phi_{a_1} \in L_p(\mathbb{R})$.

Note that $\sigma_{k,p}(a_1) := \lim_{n\to\infty} \|\nabla_1^k S_{a_1}^n \delta\|_p^{1/n}$ and

$$\sigma_{k,p}(a) := \lim_{n\to\infty} \max\{\, \|\nabla_i^k S_a^n \delta\|_p^{1/n} \;:\; i = 1, \dots, s \,\} \ge \lim_{n\to\infty} \|\nabla_1^k S_a^n \delta\|_p^{1/n}.$$

Hence, the inequality (4.4) gives rise to

$$\sigma_{k,p}(a_1) \le 2^{(1-s)/p} \sigma_{k,p}(a) \qquad \forall\, k \in \mathbb{N} \cup \{0\}.$$

Let k be a positive integer greater than $\nu_p(\phi)$. It follows from Theorem 3.3 that

$$\nu_p(\phi_{a_1}) \ge 1/p - \log_2 \sigma_{k,p}(a_1) \ge s/p - \log_2 \sigma_{k,p}(a) = \nu_p(\phi),$$

as desired. □

Combining the above lemmas and Theorem 3.5, we have the following result.

COROLLARY 4.3. *Suppose that a function ϕ is a fundamental real-valued function and satisfies the refinement equation* (1.1) *with an interpolatory refinement mask a supported on $[-3, 3]^s$. Then the inequality $\nu_\infty(\phi) \le 2$ holds true and therefore, ϕ does not belong to $C^2(\mathbb{R}^s)$.*

Proof. Let the sequence a_1 on $\mathbb{Z}$ be given in (4.1). Suppose that $\nu_\infty(\phi) > 2$. Then the mask a must satisfy the sum rules of order at least 3. Therefore, it follows from Lemma 4.1 that a_1 is an interpolatory mask. Let ϕ_{a_1} be the normalized solution of (1.1) with the mask a_1. Then by Lemma 4.2, the subdivision scheme associated with a_1 converges in the L_∞ norm which implies that the function ϕ_{a_1} is a fundamental function. From Lemma 4.2, we also have $\nu_\infty(\phi) \le \nu_\infty(\phi_{a_1})$. It follows from Theorem 3.5 that $\nu_\infty(\phi) \le \nu_\infty(\phi_{a_1}) \le 2$. This is a contradiction to our assumption $\nu_\infty(\phi) > 2$. Therefore, the inequality $\nu_\infty(\phi) \le 2$ holds true. □

Corollary 4.3 says that there is no C^2 fundamental refinable function supported on $[-3, 3]^s$. This result also implies that if a function ϕ is an orthogonal scaling function supported on $[0, 3]^s$, then $\nu_2(\phi) \le 1$ and therefore, $\phi \notin C^1(\mathbb{R}^s)$.

Let ϕ be an orthogonal scaling function with its mask supported on $[0, 2r - 1]^s$ for some positive integer r. From Corollary 2.4, we see that $S(\phi)$ can provide approximation order at most r. For this case, we shall study the upper bound of the critical exponent $\nu_p(\phi)$ for any p such that $1 \le p \le \infty$. Based on the above lemmas and Theorem 3.3, we have the following result on orthogonal scaling functions.

THEOREM 4.4. *Suppose that a function ϕ in $L_2(\mathbb{R}^s)$ is an orthogonal scaling function with its refinement mask a supported on $[0, 2r-1]^s \cap \mathbb{Z}^s$ for some positive integer r. Define a new sequence a_1 on $\mathbb{Z}$ as follows:*

$$a_1(k) := 2^{1-s} \sum_{\beta_2 \in \mathbb{Z}} \cdots \sum_{\beta_s \in \mathbb{Z}} a(k, \beta_2, \ldots, \beta_s), \qquad k \in \mathbb{Z}.$$

Let ϕ_{a_1} be the normalized solution of the refinement equation (1.1) *with the mask a_1. If the mask a satisfies the sum rules of optimal order r, then the function ϕ_{a_1} is an orthogonal scaling function with the mask a_1 satisfying*

$$\overline{\widetilde{a_1}(z)}\, \widetilde{a_1}(z) = 2\widetilde{b}_r(z), \qquad z \in \mathbb{T}.$$

If, in addition, the function ϕ belongs to $L_p(\mathbb{R}^s)$ for some p such that $1 \le p \le \infty$, then

$$\nu_p(\phi) \le \nu_p(\phi_{a_1}).$$

In particular,

$$\nu_2(\phi) \le \nu_2(\phi_{D_r}) \quad \textit{and} \quad \nu_2(\phi_{a_1}) = \nu_2(\phi_{D_r}) = \nu_\infty(\phi_{b_r})/2,$$

where ϕ_{D_r} is the Daubechies orthogonal scaling function with its mask D_r supported on $[0, 2r-1]$, and ϕ_{b_r} is the Deslauriers and Dubuc fundamental refinable function with its mask b_r supported on $[1-2r, 2r-1]$.

Proof. Let a sequence b on $\mathbb{Z}^s$ be given by its symbol

$$\widetilde{b}(z) := 2^{-s}\overline{\widetilde{a}(z)}\widetilde{a}(z), \qquad z \in \mathbb{T}^s.$$

By Lemma 2.1, the sequence b is an interpolatory refinement mask since ϕ is an orthogonal scaling function. Since the mask a satisfies the sum rules of order r, by the proof of Theorem 2.3, we see that the sequence b must satisfy the sum rules of order at least $2r$. Define a new sequence c on $\mathbb{Z}$ as in (4.1) by

$$c(k) = 2^{1-s} \sum_{\beta_2 \in \mathbb{Z}} \cdots \sum_{\beta_s \in \mathbb{Z}} b(k, \beta_2, \ldots, \beta_s), \qquad k \in \mathbb{Z}.$$

By Lemma 4.1, the sequence c must be the mask b_r since the sequence b is supported on $[1-2r, 2r-1]^s$ and satisfies the sum rules of order $2r$. Note that $\widetilde{c}(z) = 2^{1-s}\widetilde{b}(z, 1, \ldots, 1)$ and $\widetilde{a_1}(z) = 2^{1-s}\widetilde{a}(z, 1, \ldots, 1)$ for any $z \in \mathbb{T}$. Therefore,

$$\overline{\widetilde{a_1}(z)}\widetilde{a_1}(z) = 2^{2-s}\widetilde{b}(z, 1, \ldots, 1) = 2\widetilde{c}(z) = 2\widetilde{b}_r(z) \qquad \forall\, z \in \mathbb{T}.$$

Thus, the mask a_1 is the dual mask of itself for $s = 1$. By Lemma 4.2, the subdivision scheme associated with the mask a_1 converges in the L_2 norm since the function ϕ is a scaling function. Hence, the function ϕ_{a_1} is an orthogonal scaling function by Lemma 2.1. If ϕ lies in $L_p(\mathbb{R}^s)$ for some p such that $1 \le p \le \infty$, then by Lemma 4.2, we have $\nu_p(\phi) \le \nu_p(\phi_{a_1})$. Note that $\overline{\widetilde{a_1}(z)}\widetilde{a_1}(z) = 2\widetilde{b}_r(z)$ implies that $\nu_2(\phi_{a_1}) = \nu_\infty(\phi_{b_r})/2$. Since $\overline{\widetilde{D_r}(z)}\widetilde{D_r}(z) = 2b_r(z)$ for any z in $\mathbb{T}$,

$$\nu_2(\phi) \le \nu_2(\phi_{a_1}) = \nu_\infty(\phi_{b_r})/2 = \nu_2(\phi_{D_r}),$$

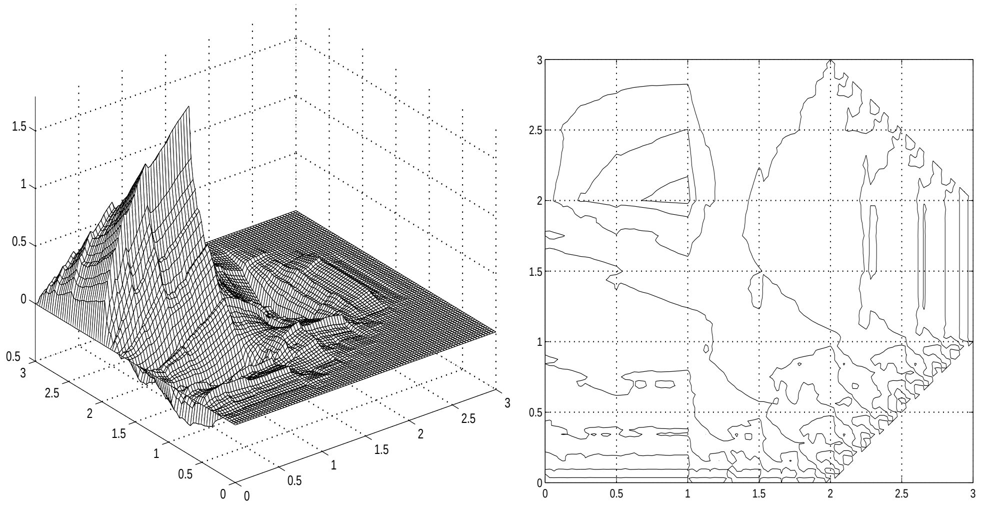


FIG. 4.1. *The graph and contour of the orthogonal scaling function ϕ_a in Example* 4.5.

which completes the proof. □

In the following, we give an example to demonstrate that when $s > 1$, such optimal orthogonal scaling functions are not unique.

Example 4.5. The mask a is supported on $[0,3]^2$ and is given by

$$\begin{pmatrix} -\frac{3}{8} - \frac{\sqrt{-10+6\sqrt{3}}}{8} + \frac{\sqrt{3}}{8} & \frac{1}{4} - \frac{\sqrt{3}}{4} & \frac{5}{8} - \frac{3\sqrt{3}}{8} + \frac{\sqrt{-10+6\sqrt{3}}}{8} & 0 \\ -\frac{1}{8} + \frac{\sqrt{3}}{8} + \frac{\sqrt{-10+6\sqrt{3}}}{8} & \frac{1}{2} & \frac{7}{8} - \frac{3\sqrt{3}}{8} - \frac{\sqrt{-10+6\sqrt{3}}}{8} & \frac{1}{4} - \frac{\sqrt{3}}{4} \\ \frac{5}{8} + \frac{\sqrt{3}}{8} + \frac{\sqrt{-10+6\sqrt{3}}}{8} & \frac{1}{2} + \frac{\sqrt{3}}{2} & \frac{1}{8} - \frac{\sqrt{-10+6\sqrt{3}}}{8} + \frac{\sqrt{3}}{8} & \frac{1}{4} - \frac{\sqrt{3}}{4} \\ \frac{3}{8} + \frac{\sqrt{3}}{8} - \frac{\sqrt{-10+6\sqrt{3}}}{8} & \frac{1}{4} + \frac{\sqrt{3}}{4} & -\frac{1}{8} + \frac{\sqrt{3}}{8} + \frac{\sqrt{-10+6\sqrt{3}}}{8} & 0 \end{pmatrix}.$$

Then the function ϕ_a is an orthogonal scaling function and the mask a satisfies the sum rules of order 2. Moreover, by calculation, we have $\nu_2(\phi_a) = 1$. Combining Theorem 4.4 and Corollary 4.3, we see that for any orthogonal scaling ϕ with its mask supported on $[0,3]^s$, the inequality $\nu_2(\phi) \le 1$ holds true. Therefore, the function ϕ_a is an optimal orthogonal scaling function in the L_2 norm sense. The graph and contour of ϕ_a are presented in Figure 4.1.

5. Optimal multivariate biorthogonal wavelets. In this section, we will demonstrate a result similar to Theorem 4.4 for the biorthogonal wavelets. Since there is a close relation between a biorthogonal wavelet and a fundamental refinable function, let us first prove the following result on fundamental refinable functions.

THEOREM 5.1. *Let ϕ be a fundamental refinable function with a finitely supported interpolatory mask a. Suppose that a is supported on $[1-2r, 2r-1]^s$ for some positive integer r and the mask a satisfies the sum rules of order $2r-1$. Let a sequence a_1 on $\mathbb{Z}$ be given by* (4.1) *and let ϕ_{a_1} be the normalized solution of the refinement equation* (1.1) *with the mask a_1. Then the function ϕ_{a_1} is a fundamental function and*

$$\nu_p(\phi) \le \nu_p(\phi_{a_1}) \qquad \forall\, 1 \le p \le \infty.$$

Moreover, if the mask a satisfies the sum rules of order $2r$, then

$$\nu_p(\phi) \le \nu_p(\phi_{b_r}) \qquad \forall\, 1 \le p \le \infty.$$

In other words, the inequality $\nu_p(\phi) \le \nu_p(\phi_{t_r})$ holds true, where t_r is the tensor product interpolatory mask given in (2.7).

Proof. By Lemma 4.1, we see that the mask a_1 is an interpolatory refinement mask. Since the function ϕ is fundamental, the shifts of ϕ are stable. By Lemma 4.2, the subdivision scheme associated with the mask a_1 converges in the L_p norm for any p such that $1 \le p \le \infty$. Hence ϕ_{a_1}, the normalized solution of the refinement equation (1.1) with the interpolatory refinement mask a_1, is continuous and therefore fundamental. It follows from Lemma 4.2 that $\nu_p(\phi) \le \nu_p(\phi_{a_1})$ for any $1 \le p \le \infty$.

If the mask a satisfies the sum rules of order $2r$, by Lemma 4.1, then the sequence a_1 must be the mask b_r. Hence, by Lemma 4.2, $\nu_p(\phi) \le \nu_p(\phi_{b_r})$ for any p such that $1 \le p \le \infty$. □

The reader is referred to [18, 21, 22, 26, 40, 41] on interpolatory subdivision schemes. In particular, a general construction of bivariate interpolatory masks g_r ($r \in \mathbb{N}$) was reported by Han and Jia in [26] with each mask g_r supported on $[1-2r, 2r-1]^2$, satisfying the optimal sum rules of order $2r$ and $\nu_2(\phi_{g_r}) = \nu_2(\phi_{b_r})$ at least for $r = 1, \ldots, 12$. Recall that by ϕ_a we denote the normalized solution of the refinement equation (1.1) with a mask a.

A similar result to Theorem 4.4 for a biorthogonal wavelet is the following.

THEOREM 5.2. *Let a function ϕ in $L_2(\mathbb{R}^s)$ be a scaling function with a refinement mask a and a function ϕ^d in $L_2(\mathbb{R}^s)$ be a dual scaling function of ϕ with a refinement mask a^d. Define two new sequences a_1 and a_1^d on $\mathbb{Z}$ as follows:*

$$a_1(k) = 2^{1-s} \sum_{\beta_2 \in \mathbb{Z}} \cdots \sum_{\beta_s \in \mathbb{Z}} a(k, \beta_2, \ldots, \beta_s), \qquad k \in \mathbb{Z},$$

and

$$a_1^d(k) = 2^{1-s} \sum_{\beta_2 \in \mathbb{Z}} \cdots \sum_{\beta_s \in \mathbb{Z}} a^d(k, \beta_2, \ldots, \beta_s), \qquad k \in \mathbb{Z}.$$

By ϕ_{a_1} and $\phi_{a_1^d}$ we denote the normalized solutions of the refinement equation (1.1) *with the masks a_1 and a_1^d, respectively. Let a sequence b on $\mathbb{Z}^s$ be given as in* (2.4) *by*

$$b(\alpha) := 2^{-s} \sum_{\beta \in \mathbb{Z}^s} \overline{a(\beta - \alpha)}\, a^d(\beta), \qquad \alpha \in \mathbb{Z}^s. \tag{5.1}$$

Suppose that the sequence b is supported on $[1-2k, 2k-1]^s \cap \mathbb{Z}^s$ for some positive integer k and b satisfies the sum rules of order $2k-1$. Then the function ϕ_{a_1} is a univariate scaling function with $\phi_{a_1^d}$ being a dual scaling function of ϕ_{a_1}. If ϕ belongs to $L_p(\mathbb{R}^s)$ and ϕ^d belongs to $L_q(\mathbb{R}^s)$ for some p, q such that $1 \le p, q \le \infty$, then $\phi_{a_1} \in L_p(\mathbb{R})$, $\phi_{a_1^d} \in L_q(\mathbb{R})$ and

$$\nu_p(\phi) \le \nu_p(\phi_{a_1}) \quad \textit{and} \quad \nu_q(\phi^d) \le \nu_q(\phi_{a_1^d}). \tag{5.2}$$

In particular, if the sequence b satisfies the sum rules of order $2k$, then

$$\overline{\widetilde{a_1}(z)}\widetilde{a_1^d}(z) = 2\widetilde{b_k}(z), \qquad z \in \mathbb{T}, \quad \textit{and} \quad \nu_q(\phi^d) \le \nu_r(\phi_{b_k}) - \nu_p(\phi),$$

where $1/r = 1/p + 1/q - 1$ *and* b_k *is the unique interpolatory mask which is supported on* $[1-2k, 2k-1]$ *and satisfies the sum rules of order* $2k$.

Proof. By Lemma 2.1, it is easily seen that the sequence b is an interpolatory mask. Let c be a sequence on $\mathbb{Z}$ given by

$$c(k) = 2^{1-s} \sum_{\beta_2 \in \mathbb{Z}} \cdots \sum_{\beta_s \in \mathbb{Z}} b(k, \beta_2, \ldots, \beta_s), \qquad k \in \mathbb{Z}.$$

It follows from Lemma 4.1 that the sequence c is an interpolatory mask since the sequence b is supported on $[1-2k, 2k-1]^s$ and satisfies the sum rules of order $2k-1$. Observe that $\widetilde{c}(z) = 2^{1-s}\widetilde{b}(z, 1, \ldots, 1)$, $\widetilde{a_1}(z) = 2^{1-s}\widetilde{a}(z, 1, \ldots, 1)$, and $\widetilde{a_1^d}(z) = 2^{1-s}\widetilde{a^d}(z, 1, \ldots, 1)$ for any z in $\mathbb{T}$. It is easy to see that

$$\overline{\widetilde{a_1}(z)}\widetilde{a_1^d}(z) = 2^{2-2s}\overline{\widetilde{a}(z, 1, \ldots, 1)}\widetilde{a^d}(z, 1, \ldots, 1) = 2\widetilde{c}(z), \qquad z \in \mathbb{T}. \tag{5.3}$$

Therefore, the masks a_1 and a_1^d must satisfy the discrete biorthogonal relation (2.5) with $s = 1$ since the sequence c is an interpolatory mask. Since both ϕ and ϕ^d belong to $L_2(\mathbb{R}^s)$ and their shifts are stable, by Lemma 4.2, the subdivision schemes associated with the masks a_1 and a_1^d converge in the L_2 norm, respectively. Thus, by Lemma 2.1, the function ϕ_{a_1} is a scaling function with $\phi_{a_1^d}$ being a dual scaling function of ϕ_{a_1}. The inequality (5.2) follows directly from Lemma 4.2.

If the sequence b satisfies the sum rules of order $2k$, by Lemma 4.1, the mask c in (5.3) must be the mask b_k. Note that $\widetilde{\nabla^{k_1} S_{a_1}^n \delta}(z) = (1-z)^{k_1}\Pi_{j=0}^{n-1}\widetilde{a_1}(z^{2^j})$. Therefore, it follows from (5.3) that for any positive integers k_1 and k_2, it is easy to verify that

$$2^n \widetilde{\nabla^{k_1+k_2} S_{b_k}^n \delta}(z) = \overline{\widetilde{\nabla^{k_1} S_{a_1}^n \delta}(z)}\widetilde{\nabla^{k_2} S_{a_1^d}^n \delta}(z), \qquad z \in \mathbb{T}.$$

Therefore, by applying Young's inequality to the above equation, we have

$$2^n \|\nabla^{k_1+k_2} S_{b_k}^n \delta\|_r \leq \|\nabla^{k_1} S_{a_1}^n \delta\|_p \|\nabla^{k_2} S_{a_1^d}^n \delta\|_q \qquad \forall\, n \in \mathbb{N},$$

where $1/r = 1/p + 1/q - 1$. This yields

$$2\sigma_{k_1+k_2,r}(b_k) \leq \sigma_{k_1,p}(a_1)\sigma_{k_2,q}(a_1^d) \qquad \forall\ k_1, k_2 \in \mathbb{N}.$$

By Theorem 3.3, we have $\nu_r(\phi_{b_k}) \geq \nu_p(\phi_{a_1}) + \nu_q(\phi_{a_1^d})$. Since $\nu_p(\phi) \leq \nu_p(\phi_{a_1})$ and $\nu_q(\phi^d) \leq \nu_q(\phi_{a_1^d})$, we conclude that $\nu_r(\phi_{b_k}) \geq \nu_p(\phi_a) + \nu_q(\phi_{a^d})$. □

COROLLARY 5.3. *Let* ϕ *be a scaling function with a refinement mask* a *supported on* $[-l, l]^s$ *for some positive integer* l *and* ϕ^d *be a dual scaling function with a refinement mask* a^d *supported on* $[1+l-2k, 2k-l-1]^s$ *for some positive integer* k. *Let the sequence* b *be given in* (5.1). *Suppose that the mask* a *satisfies the sum rules of order* m. *Then the mask* a^d *can satisfy the sum rules of order at most* $2k-m$. *Moreover, if the mask* a^d *satisfies the sum rules of order* $2k-m-1$ (*or* $2k-m$), *then the sequence* b *can satisfy the sum rules of order at least* $2k-1$ (*or* $2k$) *and the corresponding results in Theorem* 5.2 *hold true.*

Proof. This is a direct consequence of Theorems 2.3 and 5.2. □

Let us consider an example. Let ϕ be a refinable box spline function with its mask a given by its symbol

$$\widetilde{a}(z) = 2^{-s}\Pi_{j=1}^{s}(z_j^{-1} + 2 + z_j), \qquad z \in \mathbb{T}^s,$$

or

$$\widetilde{a}(z) = 2^{-1}(1 + z_1^{-1} \cdots z_s^{-1})\Pi_{j=1}^{s}(1 + z_j), \qquad z \in \mathbb{T}^s.$$

It is easy to verify that ϕ is a fundamental function with $\nu_1(\phi) = 2$, its mask a is supported on $[-1, 1]^s$, and a satisfies the sum rules of order 2. Thus, the function ϕ is a scaling function. Then Corollaries 4.3 and 5.3 imply that if a function ϕ^d is a dual scaling function of the scaling function ϕ with its mask supported on $[-2, 2]^s$, then the function ϕ^d cannot be continuous. For any dual scaling function ϕ^d of the scaling function ϕ with its mask a^d supported on $[2 - 2r, 2r - 2]^s$ for some positive integer r, by Theorem 2.3, the mask a^d can satisfy the sum rules of order at most $2r - 2$. If a^d satisfies the sum rules of order $2r - 2$, by Corollary 5.3, then we have

$$\nu_2(\phi^d) \le \nu_2(\phi_{b_r}) - \nu_1(\phi) = \nu_2(\phi_{b_r}) - 2.$$

When $s = 2$, in the next section, we shall construct a family of dual scaling functions $\phi_{\mathcal{H}_r}$ $(r \in \mathbb{N})$ of the bivariate hat function ϕ such that the dual mask $\mathcal{H}_r$ is supported on $[2-2r, 2r-2]^2$ and satisfies the sum rules of order $2r-2$. In addition, the equality $\nu_2(\phi_{\mathcal{H}_r}) = \nu_2(\phi_{b_r}) - 2$ holds true at least for $r = 3, \ldots, 12$ and each mask $\mathcal{H}_r$ is symmetric about the two coordinate axes, and the lines $x_1 = x_2$ and $x_1 = -x_2$.

6. Construction of multivariate biorthogonal wavelets. In this section, we shall present a general method to construct multivariate biorthogonal wavelets. More precisely, for any scaling function ϕ with an interpolatory refinement mask a, a general CBC algorithm is given to produce all the dual masks of the mask a. As an application of this general theory, for any bivariate fundamental mask a which is symmetric about the two coordinate axes, we construct a family of dual masks of a which satisfy any desired order of sum rules and are also symmetric about the two coordinate axes. Based on this construction, a family of optimal bivariate biorthogonal wavelets is presented. Such biorthogonal wavelets have full symmetry (i.e., they are symmetric about the x_1-axis, x_2-axis, and the lines $x_1 = x_2$ and $x_1 = -x_2$), have the optimal order of sum rules, the optimal L_2 smoothness order, and relatively small support of the dual masks.

Before proceeding further, we introduce some notation. Recall that

$$\mathbb{Z}_+^s := \{\, (\alpha_1, \ldots, \alpha_s) \in \mathbb{Z}^s \;:\; \alpha_i \ge 0 \quad \forall\, i = 1, \ldots, s \,\}.$$

For any $\mu = (\mu_1, \ldots, \mu_s) \in \mathbb{Z}^s$, we denote $|\mu| := |\mu_1| + \cdots + |\mu_s|$ and $\mu! := \mu_1! \cdots \mu_s!$ if $\mu \in \mathbb{Z}_+^s$. For any $\mu = (\mu_1, \ldots, \mu_s), \nu = (\nu_1, \ldots, \nu_s) \in \mathbb{Z}^s$, by $\nu \le \mu$ we mean $\nu_i \le \mu_i$ $\forall\, i = 1, \ldots, s$, and by $\nu < \mu$ we mean $\nu \le \mu$ and $\nu \ne \mu$.

Throughout this section, for any $\nu \in \mathbb{Z}_+^s$, by p_ν we denote the monomial $(\cdot)^\nu$ and

$$\langle \lambda, p_\nu \rangle := \sum_{\alpha \in \mathbb{Z}^s} \lambda(\alpha) p_\nu(\alpha) = \sum_{\alpha \in \mathbb{Z}^s} \lambda(\alpha)\alpha^\nu, \qquad \lambda \in \ell_0(\mathbb{Z}^s).$$

THEOREM 6.1. *Let a sequence a on $\mathbb{Z}^s$ satisfy $\sum_{\beta \in \mathbb{Z}^s} a(\beta) = 2^s$. Suppose that a sequence a^d on $\mathbb{Z}^s$ is a dual mask of a that satisfies the following discrete biorthogonal relation:*

$$\sum_{\beta \in \mathbb{Z}^s} \overline{a(\beta - 2\alpha)} a^d(\beta) = 2^s \delta(\alpha) \qquad \forall\, \alpha \in \mathbb{Z}^s. \tag{6.1}$$

If the sequence a^d satisfies the sum rules of order k for some positive integer k, then for any $\mu \in \mathbb{Z}_+^s$ such that $|\mu| < k$, the value $h_\mu := 2^{-s}\langle a^d, p_\mu\rangle$ is uniquely determined by the sequence a. More precisely, h_μ is given by the following recursive relation:

$$h_\mu = \delta(\mu) - 2^{-s} \sum_{0 \le \nu < \mu} (-1)^{|\mu-\nu|} \frac{\mu!}{\nu!(\mu-\nu)!} \overline{\langle a, p_{\mu-\nu}\rangle}\, h_\nu, \qquad |\mu| < k, \mu \in \mathbb{Z}_+^s. \tag{6.2}$$

Proof. Recall that Ω is the set of the vertices of the unit cube $[0,1]^s$. By the definition of the sum rules (1.4), we observe that the sequence a^d satisfies the sum rules of order k if and only if

$$\sum_{\beta \in \mathbb{Z}^s} a^d(2\beta+\varepsilon)(2\beta+\varepsilon)^\nu = 2^{-s}\langle a^d,\, p_\nu\rangle = h_\nu \qquad \forall\, \varepsilon \in \Omega,\ |\nu| < k,\ \nu \in \mathbb{Z}_+^s. \tag{6.3}$$

From (6.1), we get for any $\mu \in \mathbb{Z}_+^s$,

$$\begin{aligned} 2^s\delta(\mu) &= \sum_{\alpha \in \mathbb{Z}^s} \sum_{\beta \in \mathbb{Z}^s} \overline{a(\beta - 2\alpha)} a^d(\beta)(2\alpha)^\mu \\ &= \sum_{\varepsilon \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \sum_{\beta \in \mathbb{Z}^s} \overline{a(2\beta+\varepsilon-2\alpha)} a^d(2\beta+\varepsilon)(2\alpha)^\mu. \end{aligned}$$

On the other hand, we have

$$(2\alpha)^\mu = \big((2\beta+\varepsilon) - (2\beta+\varepsilon-2\alpha)\big)^\mu = \sum_{0 \le \nu \le \mu} c_{\nu,\mu}(2\beta+\varepsilon-2\alpha)^{\mu-\nu}(2\beta+\varepsilon)^\nu,$$

where $c_{\nu,\mu} := (-1)^{|\mu-\nu|}\mu!/(\nu!(\mu-\nu)!)$ and recall that by $\nu \le \mu$ we mean $\nu_i \le \mu_i$ $\forall\, i = 1, \dots, s$. Hence, for any $\mu \in \mathbb{Z}_+^s$, we deduce that

$$\begin{aligned} &2^s\delta(\mu) \\ &\quad = \sum_{0 \le \nu \le \mu} c_{\nu,\mu} \sum_{\varepsilon \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \sum_{\beta \in \mathbb{Z}^s} \overline{a(2\beta+\varepsilon-2\alpha)}(2\beta+\varepsilon-2\alpha)^{\mu-\nu} a^d(2\beta+\varepsilon)(2\beta+\varepsilon)^\nu \\ &\quad = \sum_{0 \le \nu \le \mu} c_{\nu,\mu} \sum_{\varepsilon \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \overline{a(2\alpha+\varepsilon)}(2\alpha+\varepsilon)^{\mu-\nu} \sum_{\beta \in \mathbb{Z}^s} a^d(2\beta+\varepsilon)(2\beta+\varepsilon)^\nu. \end{aligned}$$

Since $\sum_{\beta \in \mathbb{Z}^s} a(\beta) = 2^s$, we have $\langle a,\, p_0\rangle = 2^s$. Note that $c_{\mu,\mu} = 1$ for any $\mu \in \mathbb{Z}_+^s$. From (6.3), we conclude that

$$\begin{aligned} 2^s\delta(\mu) &= \sum_{0 \le \nu \le \mu} c_{\nu,\mu} \sum_{\varepsilon \in \Omega} \sum_{\alpha \in \mathbb{Z}^s} \overline{a(2\alpha+\varepsilon)}(2\alpha+\varepsilon)^{\mu-\nu}\, h_\nu \\ &= \sum_{0 \le \nu \le \mu} c_{\nu,\mu} \overline{\langle a,\, p_{\mu-\nu}\rangle}\, h_\nu = 2^s h_\mu + \sum_{0 \le \nu < \mu} c_{\nu,\mu} \overline{\langle a,\, p_{\mu-\nu}\rangle}\, h_\nu \end{aligned}$$

from which (6.2) can be easily derived. □

By definition, the value $\langle a^d,\, p_\mu\rangle$ in Theorem 6.1 is totally determined by the sequence a^d. But Theorem 6.1 says that if the sequence a^d is a dual mask of the mask a and the sequence a^d satisfies the sum rules of order k, then for any $\mu \in \mathbb{Z}_+^s$ such that $|\mu| < k$, the value $\langle a^d,\, p_\mu\rangle$ is uniquely determined by the sequence a instead of the sequence a^d. Therefore, by the above Theorem 6.1, if a sequence a on $\mathbb{Z}^s$ satisfies $\sum_{\beta \in \mathbb{Z}^s} a(\beta) = 2^s$, then we can define a sequence h^a on $\mathbb{Z}_+^s$ as follows:

$$h^a(\mu) = \delta(\mu) - 2^{-s} \sum_{0 \le \nu < \mu} (-1)^{|\mu-\nu|} \frac{\mu!}{\nu!(\mu-\nu)!} \overline{\langle a, p_{\mu-\nu}\rangle}\, h^a(\nu), \qquad \mu \in \mathbb{Z}_+^s. \tag{6.4}$$

An important consequence of Theorem 6.1 is that it allows us to propose a general method to construct a dual mask satisfying the sum rules of arbitrary order for any given interpolatory refinement mask. Since, in the following method, we obtain the dual masks a^d by constructing each coset $a^d(2\beta+\varepsilon), \beta \in \mathbb{Z}^s$ separately, we call this method the CBC algorithm.

CBC ALGORITHM.

(1) *Given a sequence a on $\mathbb{Z}^s$ such that a satisfies the following conditions:*

$$(6.5) \quad \sum_{\beta\in\mathbb{Z}^s} a(\beta) = 2^s, \quad a(0)=1, \quad \textit{and} \quad a(2\beta)=0 \qquad \forall\, \beta \in \mathbb{Z}^s\backslash\{0\};$$

(2) *let k be any fixed positive integer;*

(3) *calculate $h^a(\mu)$ as in (6.4) for $\mu \in \mathbb{Z}_+^s$ such that $|\mu| < k$;*

(4) *let Ω be the set of vertices of $[0,1]^s$. For each $\varepsilon \in \Omega\backslash\{0\}$, choose an appropriate subset E_ε of $\mathbb{Z}^s$ such that the following linear system*

$$(6.6) \qquad \sum_{\beta\in E_\varepsilon} b_{\varepsilon,\beta}(2\beta+\varepsilon)^\mu = h^a(\mu), \qquad \mu\in\mathbb{Z}_+^s,\ |\mu|<k$$

has at least one solution for $\{b_{\varepsilon,\beta} \,:\, \beta \in E_\varepsilon\}$;

(5) *construct the mask a^d coset by coset as follows: for each $\varepsilon \in \Omega\backslash\{0\}$,*

$$a^d(2\beta+\varepsilon) = b_{\varepsilon,\beta},\ \beta\in E_\varepsilon \quad \textit{and} \quad a^d(2\beta+\varepsilon)=0, \quad \beta\in\mathbb{Z}^s\backslash E_\varepsilon$$

and

$$(6.7) \quad a^d(2\beta) = 2^s\delta(\beta) - \sum_{\varepsilon\in\Omega\backslash\{0\}}\sum_{\alpha\in\mathbb{Z}^s} \overline{a(2\alpha-2\beta+\varepsilon)}\, a^d(2\alpha+\varepsilon),\ \beta\in\mathbb{Z}^s;$$

(6) *then the mask a^d is a dual mask of the given interpolatory mask a and satisfies the sum rules of order k.*

Proof. It is easy to verify that if the sequence a is an interpolatory mask, then the dual relation (6.1) is equivalent to (6.7). Therefore, the mask a^d is a dual mask of the given mask a. On the other hand, (6.6) can be rewritten as

$$(6.8) \qquad \sum_{\beta\in\mathbb{Z}^s} a^d(2\beta+\varepsilon)(2\beta+\varepsilon)^\mu = h^a(\mu), \qquad \varepsilon\in\Omega\backslash\{0\},\ |\mu|<k,\ \mu\in\mathbb{Z}_+^s.$$

By the definition of sum rules, to verify that the sequence a^d satisfies the sum rules of order k, it suffices to demonstrate that

$$(6.9) \qquad \sum_{\beta\in\mathbb{Z}^s} a^d(2\beta)(2\beta)^\mu = h^a(\mu) \qquad \forall\,|\mu|<k,\ \mu\in\mathbb{Z}_+^s.$$

As in the proof of Theorem 6.1, from (6.7), we have

$$\begin{aligned}&\sum_{\beta\in\mathbb{Z}^s} a^d(2\beta)(2\beta)^\mu\\ &= 2^s\delta(\mu) - \sum_{0\le\nu\le\mu} c_{\nu,\mu} \sum_{\varepsilon\in\Omega\backslash\{0\}}\sum_{\alpha\in\mathbb{Z}^s} \overline{a(2\alpha+\varepsilon)}(2\alpha+\varepsilon)^{\mu-\nu} \sum_{\beta\in\mathbb{Z}^s} a^d(2\beta+\varepsilon)(2\beta+\varepsilon)^\nu,\end{aligned}$$

where $c_{\nu,\mu} := (-1)^{|\mu-\nu|}\mu!/(\nu!(\mu-\nu)!)$. Since the sequence a is an interpolatory mask, it is easily seen that

$$\sum_{\varepsilon\in\Omega\backslash\{0\}}\sum_{\alpha\in\mathbb{Z}^s}\overline{a(2\alpha+\varepsilon)}(2\alpha+\varepsilon)^{\mu-\nu} = \overline{\langle a,\ p_{\mu-\nu}\rangle} - \delta(\mu-\nu).$$

Therefore, it follows from (6.8) that for any $\mu \in \mathbb{Z}_+^s$ such that $|\mu| < k$,

$$\begin{aligned}\sum_{\beta\in\mathbb{Z}^s} a^d(2\beta)(2\beta)^\mu &= 2^s\delta(\mu) - \sum_{0\le\nu\le\mu} c_{\nu,\mu}\left(\overline{\langle a,\ p_{\mu-\nu}\rangle} - \delta(\mu-\nu)\right) h^a(\nu)\\ &= (1-2^s)h^a(\mu) + 2^s\delta(\mu) - \sum_{0\le\nu<\mu} c_{\nu,\mu}\overline{\langle a,\ p_{\mu-\nu}\rangle}\, h^a(\nu)\\ &= h^a(\mu),\end{aligned}$$

where in the last equality we used (6.4) for $h^a(\mu)$. This completes the proof. □

It is evident that the above CBC algorithm can produce all the dual masks for any given interpolatory mask. In general, if the set E_ε is large enough, the equation in step (4) must have at least one solution. We point out that based on Theorem 6.1, the CBC algorithm can be generalized to the general case. In a forthcoming paper, we shall propose a similar CBC algorithm such that for any given scaling function with a mask a, we can construct a dual mask of the mask a which can satisfy the sum rules of arbitrary order. Based on the work [26], here we present a concrete way to implement the above general CBC algorithm in the bivariate case. By $\#E$ we denote the cardinality of a set E. Let us cite a result from [26].

LEMMA 6.2 (see [26, Lemma 4.1]). *Let $l_1,\dots,l_r$ be distinct parallel lines in $\mathbb{R}^2$ and let E be a subset of $l_1\cup\cdots\cup l_r$ such that $\#(E\cap l_j) = j$ for each $j = 1,\dots,r$. Suppose p is a polynomial in two variables of (total) degree at most $r-1$. If p vanishes on E, then p vanishes everywhere. Consequently, the square matrix $(t_1^{\nu_1}t_2^{\nu_2})_{(t_1,t_2)\in E, 0\le\nu_1+\nu_2\le r-1}$ is nonsingular.*

Now for any bivariate interpolatory mask a which is symmetric about the two coordinate axes, the following algorithm provides us with a method to construct a dual mask which satisfies the sum rules of arbitrary order.

TCBC ALGORITHM.

(1) *Let a bivariate mask a satisfy $\sum_{\beta\in\mathbb{Z}^2} a(\beta) = 4, a(0) = 1,$ and $a(2\beta) = 0$ $\forall\, \beta \in \mathbb{Z}^2\backslash\{0\}$ with symmetry about the two coordinate axes, i.e.,*

$$a(\beta_1,\beta_2) = a(-\beta_1,\beta_2) = a(\beta_1,-\beta_2) = a(-\beta_1,-\beta_2); \tag{6.10}$$

(2) *let k be any fixed positive integer;*

(3) *calculate $h^a(2\mu)$ as in (6.4) for $\mu \in \mathbb{Z}_+^2$ such that $|\mu| < k$;*

(4) *let the set $E := \{(\beta_1,\beta_2) \in \mathbb{Z}^2\ :\ \beta_1 \ge 0,\ \beta_2 \ge 0$ and $\beta_1+\beta_2 < k\}$;*

(5) *let $\Omega' := \{(1,0),(0,1),(1,1)\}$. For each $\varepsilon \in \Omega'$, there is a unique solution for $\{b_{\varepsilon,\beta}, \beta \in E\}$ to the following linear system:*

$$\sum_{\beta\in E} b_{\varepsilon,\beta}(2\beta+\varepsilon)^{2\mu} = h^a(2\mu)/4, \qquad |\mu| < k, \mu \in \mathbb{Z}_+^2;$$

(6) *for each $(\varepsilon_1,\varepsilon_2) \in \Omega'$, set $a^d(2\beta_1+\varepsilon_1, 2\beta_2+\varepsilon_2) = 0\ \forall\,(\beta_1,\beta_2) \in \mathbb{Z}_+^2\backslash E$, and for any $(\beta_1,\beta_2) \in E$,*

$$a^d(2\beta_1+\varepsilon_1, 2\beta_2+\varepsilon_2) = \big(1+\delta(2\beta_1+\varepsilon_1)\big)\big(1+\delta(2\beta_2+\varepsilon_2)\big)b_{(\varepsilon_1,\varepsilon_2),(\beta_1,\beta_2)};$$

(7) *for each* $\varepsilon \in \Omega'$*, complete each coset* $a(2\beta+\varepsilon), \beta \in \mathbb{Z}^2$*, by symmetry as in* (6.10) *and set*

$$a^d(2\beta) := 4\delta(\beta) - \sum_{\varepsilon\in\Omega'} \sum_{\alpha\in\mathbb{Z}^2} \overline{a(2\alpha-2\beta+\varepsilon)}a^d(2\alpha+\varepsilon), \quad \beta \in \mathbb{Z}^2;$$

(8) *then the mask* a^d *is a dual mask of the given mask* a*, satisfies the sum rules of order* $2k$*, and it is symmetric about the two coordinate axes.*

The above algorithm is called the TCBC algorithm since we choose a special triangle subset E of $\mathbb{Z}^2$ it. The existence and uniqueness of the solution in step (5) of the TCBC algorithm are guaranteed by Lemma 6.2. The claim that the mask a^d satisfies the sum rules of order $2k$ follows from the fact that if the sequence a is symmetric about the two coordinate axes, then $\langle a, p_{(\nu_1,\nu_2)}\rangle = 0$ for any $(\nu_1,\nu_2) \in \mathbb{Z}_+^2$ with either ν_1 or ν_2 being an odd integer. We mention that if, in the TCBC algorithm, the mask a is also symmetric about the lines $x_1 = x_2$ and $x_1 = -x_2$, then the resulting dual mask also has such properties. For this case, in step (5) of the TCBC algorithm, we need to deal only with the coset of a^d at $2\beta+\varepsilon, \beta \in \mathbb{Z}^2$, for $\varepsilon \in \{(1,0),(1,1)\}$. The coset of the mask a^d at $2\beta+(0,1), \beta\in\mathbb{Z}^2$, is obtained by symmetry.

Let us illustrate the above general theory by giving an example. Let φ_h be the bivariate hat function with its mask a_h supported on $[-1,1]^2 \cap \mathbb{Z}^2$ and given by

$$\begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{pmatrix}. \tag{6.11}$$

An easy calculation gives us

$$\langle a, p_{(\mu_1,\mu_2)}\rangle = \begin{cases} \big(1+\delta(\mu_1)\big)\big(1+\delta(\mu_2)\big), & (\mu_1,\mu_2) \in 2\mathbb{Z}_+^2; \\ 0 & \text{otherwise.} \end{cases}$$

Let H_r denote the dual mask of the mask a_h derived by the TCBC algorithm such that H_r satisfies the sum rules of order $2r-2$.

From the TCBC algorithm, it is easily seen that

$$\operatorname{supp} H_r \subseteq \{\, (\beta_1,\beta_2) \in \mathbb{Z}^2 \;:\; |\beta_1| \le 2r-2,\, |\beta_2| \le 2r-2,\, |\beta_1|+|\beta_2| \le 2r \,\}$$

and it is symmetric about the x_1-axis, x_2-axis, and the lines $x_1 = x_2$ and $x_1 = -x_2$.

Let us give an example of the dual mask H_r for $r = 3$ in the following.

Example 6.3. The mask H_3 is supported on $[-4,4]^2$ and is given by

$$\begin{pmatrix}
0 & 0 & \frac{3}{256} & 0 & \frac{9}{128} & 0 & \frac{3}{256} & 0 & 0 \\
0 & 0 & 0 & -\frac{3}{64} & -\frac{3}{32} & -\frac{3}{64} & 0 & 0 & 0 \\
\frac{3}{256} & 0 & -\frac{1}{32} & -\frac{1}{32} & -\frac{51}{128} & -\frac{1}{32} & -\frac{1}{32} & 0 & \frac{3}{256} \\
0 & -\frac{3}{64} & -\frac{1}{32} & \frac{11}{32} & \frac{21}{32} & \frac{11}{32} & -\frac{1}{32} & -\frac{3}{64} & 0 \\
\frac{9}{128} & -\frac{3}{32} & -\frac{51}{128} & \frac{21}{32} & \frac{75}{32} & \frac{21}{32} & -\frac{51}{128} & -\frac{3}{32} & \frac{9}{128} \\
0 & -\frac{3}{64} & -\frac{1}{32} & \frac{11}{32} & \frac{21}{32} & \frac{11}{32} & -\frac{1}{32} & -\frac{3}{64} & 0 \\
\frac{3}{256} & 0 & -\frac{1}{32} & -\frac{1}{32} & -\frac{51}{128} & -\frac{1}{32} & -\frac{1}{32} & 0 & \frac{3}{256} \\
0 & 0 & 0 & -\frac{3}{64} & -\frac{3}{32} & -\frac{3}{64} & 0 & 0 & 0 \\
0 & 0 & \frac{3}{256} & 0 & \frac{9}{128} & 0 & \frac{3}{256} & 0 & 0
\end{pmatrix}.$$

It is a dual mask of the mask a_h and satisfies the sum rules of order 4. By calculation, we have $\nu_2(\phi_{H_3}) \approx 0.42927$. Therefore, by Theorem 2.3 and Corollary 5.3, the dual scaling function ϕ_{H_3} attains the optimal sum rules but does not attain the optimal L_2 smoothness order since $\nu_2(\phi_{H_3}) < \nu_2(\phi_{b_3}) - \nu_1(\varphi_h) \approx 1.17513$.

In the rest of this section, we shall modify the above TCBC algorithm to construct a new family of optimal biorthogonal wavelets by shrinking the support of each H_r. Since the mask a_h has full symmetry, we need to deal only with $\varepsilon \in \{(1,0),(1,1)\}$ in step (5) of the TCBC algorithm. The only part we need to modify in the TCBC algorithm is steps (5) and (6). All the other steps are the same. Throughout the rest of this section, the mask a in the TCBC algorithm is assumed to be a_h given in (6.11).

Let the set E be given in step (4) of the TCBC algorithm and let $b_\beta, \beta \in E$, be the unique solution of the following linear system:

$$\sum_{\beta\in E} b_\beta\big(2\beta+(1,1)\big)^{2\mu} = h^a(2\mu)/4, \qquad |\mu|<k, \mu\in\mathbb{Z}_+^2.$$

Set $a^d(2\beta+(1,1)) = b_\beta, \beta \in E$, and $a^d(2\beta+(1,1)) = 0, \beta \in \mathbb{Z}_+^2\backslash E$. Take F to be the following set:

$$F := \{\,(\beta_1,\beta_2)\in\mathbb{Z}_+^2 \ :\ \beta_1+\beta_2 = k \quad\text{and}\quad \beta_2>0\,\}.$$

Now we set $a^d(2\beta+(1,0)) = 0$ for any $\beta \in \mathbb{Z}_+^2\backslash(E\cup F)$ and

$$a^d(2\beta_1+1,2\beta_2) = \big(1+\delta(\beta_2)\big)c_{(\beta_1,\beta_2)}, \qquad (\beta_1,\beta_2)\in E\cup F,$$

with some yet-to-be-determined parameters $c_\beta, \beta \in E\cup F$.

This extra freedom $c_\beta, \beta \in F$, given by F will be used to reduce the support of the mask a^d at the coset $(0,0)$ constructed in step (7) of the TCBC algorithm. More precisely, we try to adjust the coefficients of H_{k-1} on the set $\{(\beta_1,\beta_2)\in\mathbb{Z}^2 : \beta_1+\beta_2 = 2k-2\}$ to be zero. By using symmetry, after a simple calculation, it is easily seen that this restriction is equivalent to the following linear system:

$$c_{(\beta_1,\beta_2)} + c_{(\beta_2-1,\beta_1+1)} + b_{(\beta_1,\beta_2-1)}/2 = 0 \qquad \forall\,(\beta_1,\beta_2)\in F.$$

By simply setting $c_{(\beta_1,\beta_2)} = 0$ for any $(\beta_1,\beta_2)\in F$ such that $k/2 \le \beta_1 < k$, the above linear system has a unique solution $c_\beta, \beta \in F$. Now the following linear system:

$$\sum_{\beta\in E} c_\beta\big(2\beta+(1,0)\big)^{2\mu} = h^a(2\mu)/4 - \sum_{\beta\in F} c_\beta\big(2\beta+(1,0)\big)^{2\mu}, \qquad |\mu|<k, \mu\in\mathbb{Z}_+^2,$$

has a unique solution for $c_\beta, \beta \in E$, by Lemma 6.2.

Set $a^d(2\beta_1,2\beta_2+1) = a^d(2\beta_2+1,2\beta_1), (\beta_1,\beta_2)\in\mathbb{Z}_+^2$. By the TCBC algorithm, we have a dual mask a^d of the mask a_h such that a^d satisfies the sum rules of order $2k$. We shall use $\mathcal{H}_r$ to denote the dual mask of the mask a_h derived from the above modified TCBC algorithm such that $\mathcal{H}_r$ satisfies the sum rules of order $2r-2$. For each positive integer r, by G_r we denote the following set:

$$G_r := \{(\alpha_1,\alpha_2)\in\mathbb{Z}^2 \ :\ |\alpha_1|+|\alpha_2| = 2r-1 \text{ and either } |\alpha_1| \text{ or } |\alpha_2| \\ \text{is an even number less than } r-1\}.$$

To sum up and restate the above construction of the dual masks $\mathcal{H}_r$ of the mask a_h, we have the following theorem.

Theorem 6.4. *Let r be a positive integer greater than two. Then there exists a unique refinement mask $\mathcal{H}_r$ satisfying the following conditions:*

(1) *$\mathcal{H}_r$ is supported on*

$$\big\{ (\alpha_1, \alpha_2) \in \mathbb{Z}^2 \ : \ |\alpha_1| + |\alpha_2| \le 2r-1, \ \max\{ |\alpha_1|, |\alpha_2| \} \le 2r-2 \big\} \backslash G_r;$$

(2) *$\mathcal{H}_r$ is symmetric about the two coordinate axes, the lines $x_1 = x_2, x_1 = -x_2$;*
(3) *$\mathcal{H}_r$ satisfies the sum rules of order $2r-2$;*
(4) *$\mathcal{H}_r$ and a_h (the mask a_h is given in (6.11)) satisfy the dual relation (6.1).*

Remark 6.5. The set G_r appears strange. The reason is that, in our modified TCBC algorithm, we set $c_{(\beta_1,\beta_2)} = 0$ for any $(\beta_1, \beta_2) \in F$ such that $(r-1)/2 \le \beta_1 < r-1$. Note that both H_r and $\mathcal{H}_r$ are symmetric about the x_1-axis, x_2-axis, and the lines $x_1 = x_2$ and $x_1 = -x_2$. Let a be a multivariate interpolatory mask such that a satisfies the sum rules of order k. For any positive integer r, by convolution, it is easy to obtain a new interpolatory mask b such that $\widetilde{b}(z) = \big(\widetilde{a}(z)\big)^r \widetilde{c_r}(z)$, $z \in \mathbb{T}^s$, where $\widetilde{c_r}(z)$ can be explicitly expressed by using $\widetilde{a}(z)$. Such an interpolatory mask b satisfies the sum rules of order rk by Theorem 2.2. See Proposition 3.7 in Han [24] for a detailed discussion on construction of biorthogonal wavelets using this convolution method. Such a method was further discussed by Ji, Riemenschneider, and Shen [27]. The TCBC algorithm proposed in this paper can be generalized to the general case, and it has many advantages over the convolution method. We shall illustrate the advantages of our CBC and TCBC algorithms over the convolution method and other known methods in the literature on construction of biorthogonal wavelets elsewhere.

Let us provide detail in the following for the masks $\mathcal{H}_3$ and $\mathcal{H}_4$.

Example 6.6. The mask $\mathcal{H}_3$ is supported on $[-4,4]^2$ and is given by

$$\begin{pmatrix}
0 & 0 & 0 & \frac{3}{128} & \frac{3}{64} & \frac{3}{128} & 0 & 0 & 0 \\
0 & 0 & 0 & -\frac{3}{64} & -\frac{3}{32} & -\frac{3}{64} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{16} & -\frac{1}{8} & -\frac{3}{8} & -\frac{1}{8} & \frac{1}{16} & 0 & 0 \\
\frac{3}{128} & -\frac{3}{64} & -\frac{1}{8} & \frac{11}{32} & \frac{51}{64} & \frac{11}{32} & -\frac{1}{8} & -\frac{3}{64} & \frac{3}{128} \\
\frac{3}{64} & -\frac{3}{32} & -\frac{3}{8} & \frac{51}{64} & \frac{33}{16} & \frac{51}{64} & -\frac{3}{8} & -\frac{3}{32} & \frac{3}{64} \\
\frac{3}{128} & -\frac{3}{64} & -\frac{1}{8} & \frac{11}{32} & \frac{51}{64} & \frac{11}{32} & -\frac{1}{8} & -\frac{3}{64} & \frac{3}{128} \\
0 & 0 & \frac{1}{16} & -\frac{1}{8} & -\frac{3}{8} & -\frac{1}{8} & \frac{1}{16} & 0 & 0 \\
0 & 0 & 0 & -\frac{3}{64} & -\frac{3}{32} & -\frac{3}{64} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{3}{128} & \frac{3}{64} & \frac{3}{128} & 0 & 0 & 0
\end{pmatrix}.$$

Then the mask $\mathcal{H}_3$ satisfies the sum rules of order 4 and $\phi_{\mathcal{H}_3}$ is a dual scaling function of φ_h with $\nu_2(\phi_{\mathcal{H}_3}) \approx 1.17513$. Thus, the function $\phi_{\mathcal{H}_3}$ is an optimal dual scaling function of the function φ_h in the L_2 norm sense since $\nu_2(\phi_{\mathcal{H}_3}) \approx \nu_2(\phi_{b_3}) - \nu_1(\varphi_h)$.

The graphs and contours of the scaling function φ_h and the dual scaling function $\phi_{\mathcal{H}_3}$ with their associated wavelets are given in Figures 6.2–6.4.

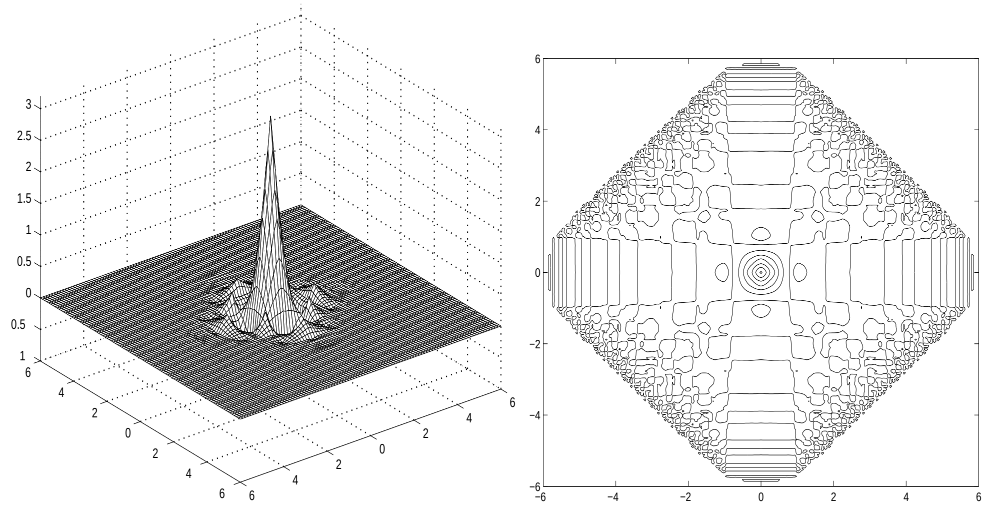


FIG. 6.1. *The graph and contour of the function* $\phi_{\mathcal{H}_4}$.

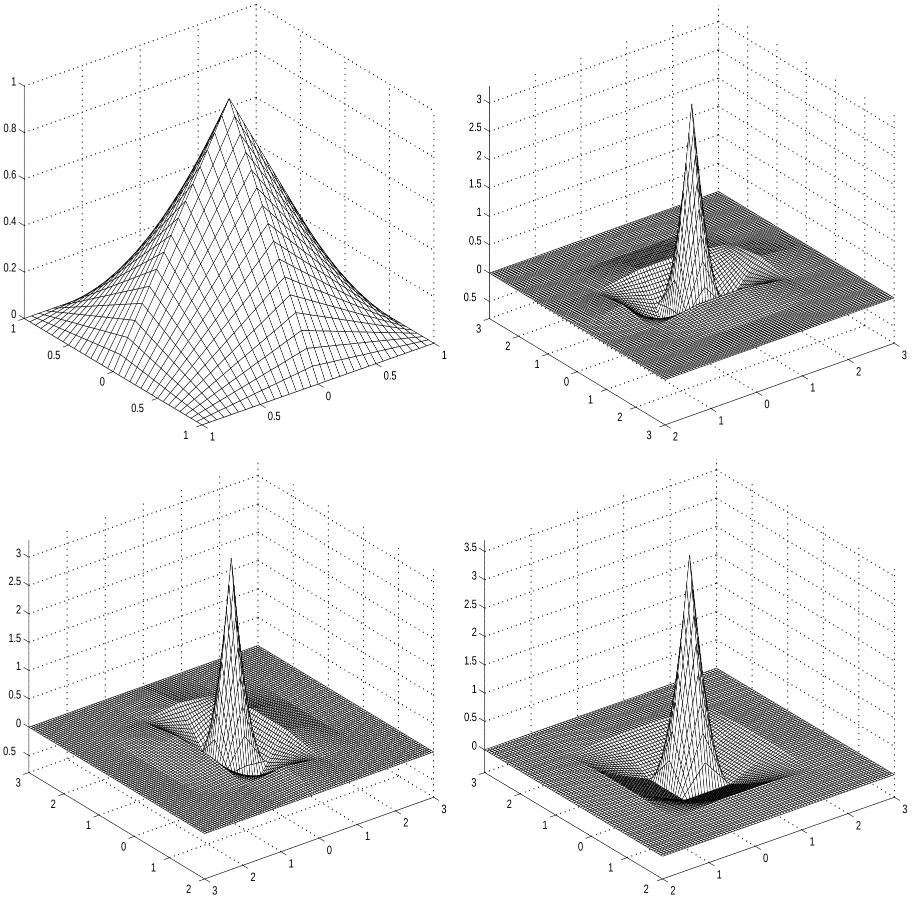


FIG. 6.2. (a) *is the scaling function* φ_h *and* (b), (c), *and* (d) *are the associated three wavelets* ψ_1, ψ_2, *and* ψ_3 *in Example* 6.6.

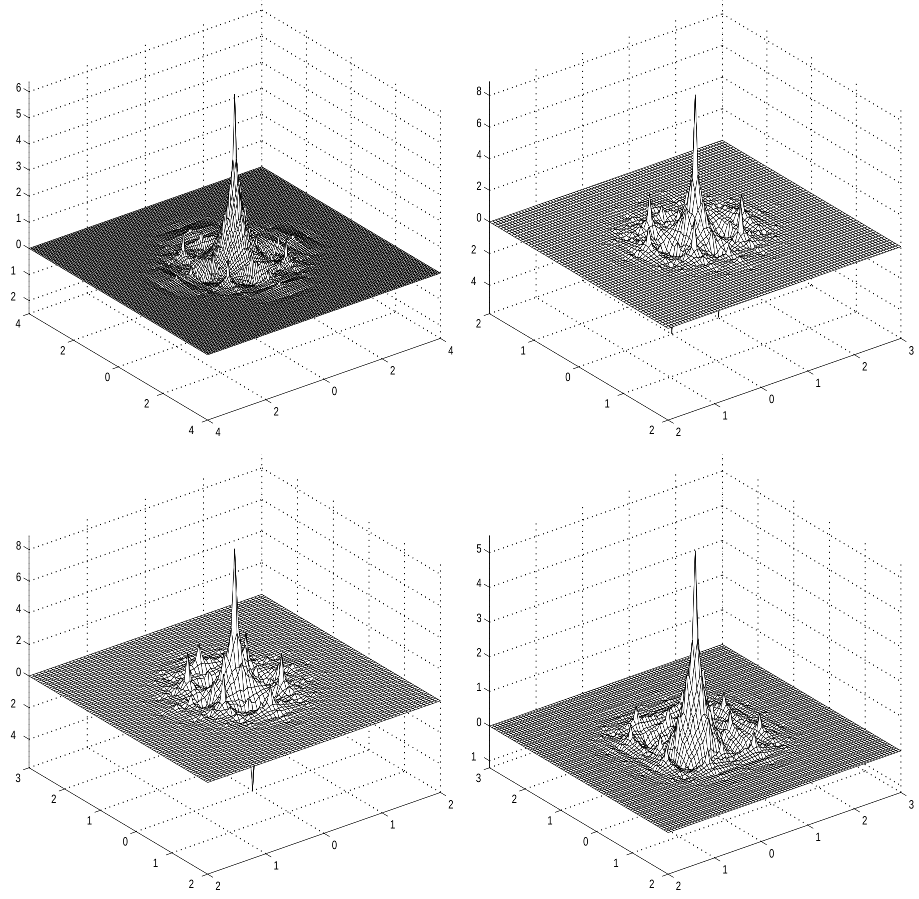


FIG. 6.3. (a) *is the dual scaling function* $\phi_{\mathcal{H}_3}$ *and* (b), (c), *and* (d) *are the associated three dual wavelets* ψ_1^d, ψ_2^d, *and* ψ_3^d *in Example* 6.6.

Example 6.7. The mask $\mathcal{H}_4$ is supported on $[-6,6]^2$ and the part of $\mathcal{H}_4$ in the first quadrant is supported on $[0,6]^2$ and is given by

$$\begin{pmatrix}
-\frac{5}{512} & -\frac{5}{1024} & 0 & 0 & 0 & 0 & 0 \\
\frac{5}{256} & \frac{5}{512} & 0 & 0 & 0 & 0 & 0 \\
\frac{83}{1024} & \frac{145}{4096} & -\frac{15}{2048} & -\frac{9}{4096} & 0 & 0 & 0 \\
-\frac{363}{2048} & -\frac{87}{1024} & \frac{15}{1024} & \frac{9}{1024} & -\frac{9}{4096} & 0 & 0 \\
-\frac{359}{1024} & -\frac{69}{512} & \frac{1}{16} & \frac{15}{1024} & -\frac{15}{2048} & 0 & 0 \\
\frac{1723}{2048} & \frac{401}{1024} & -\frac{69}{512} & -\frac{87}{1024} & \frac{145}{4096} & \frac{5}{512} & -\frac{5}{1024} \\
\frac{493}{256} & \frac{1723}{2048} & -\frac{359}{1024} & -\frac{363}{2048} & \frac{83}{1024} & \frac{5}{256} & -\frac{5}{512}
\end{pmatrix}$$

with the number $\frac{493}{256}$ at the bottom-left as $\mathcal{H}_4(0,0)$. Since $\mathcal{H}_4$ is symmetric about the coordinate axes, other parts of $\mathcal{H}_4$ are obtained by symmetry as in (6.10). By calculation, we have $\nu_2(\phi_{\mathcal{H}_4}) \approx 1.79313$ and the mask $\mathcal{H}_4$ satisfies the sum rules of order 6. Thus, the function $\phi_{\mathcal{H}_4}$ is an optimal dual scaling function of φ_h in the L_2

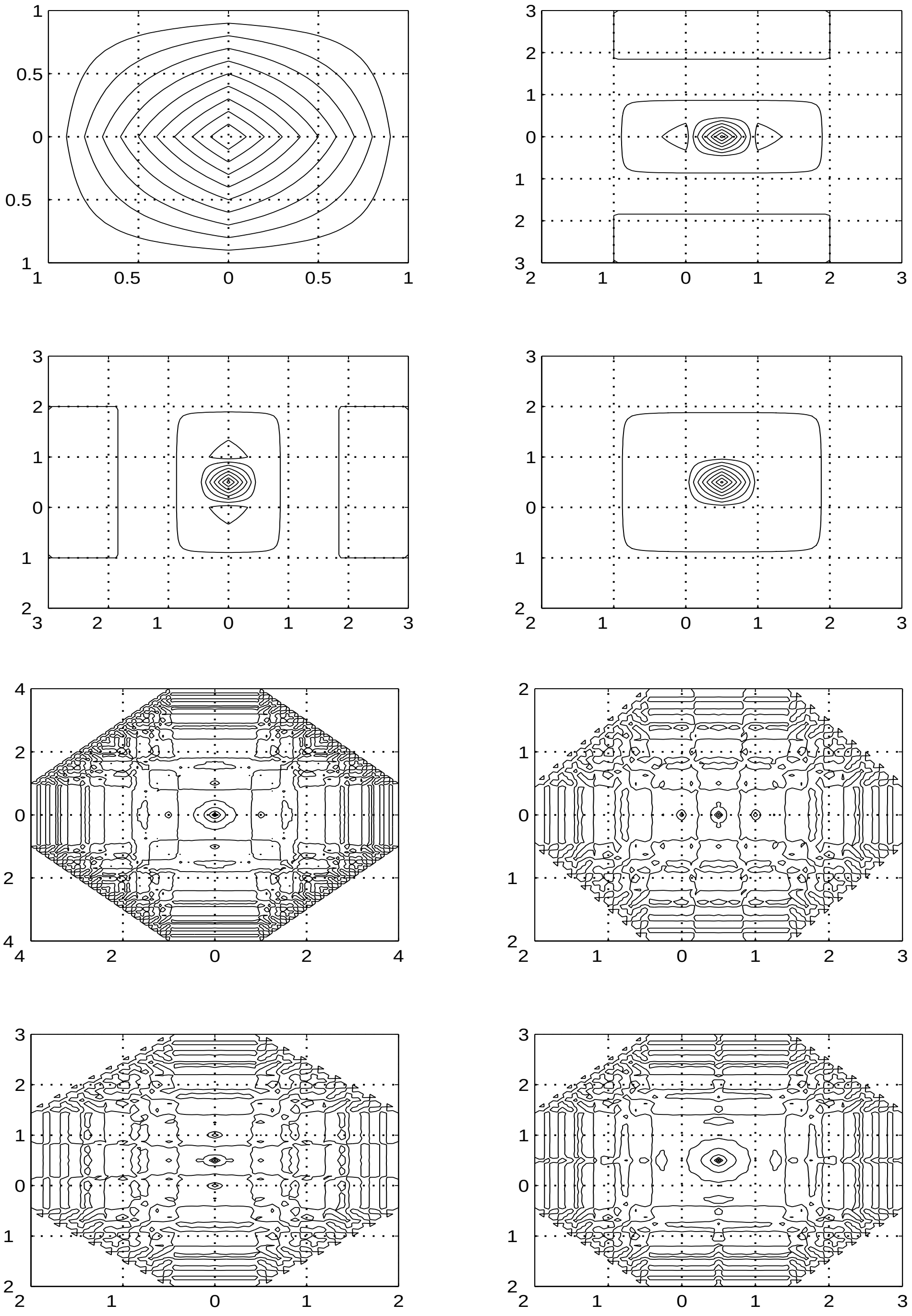


Fig. 6.4. *The contours of the scaling function* φ_h*, its dual scaling function* $\phi_{\mathcal{H}_3}$*, and the associated wavelets and dual wavelets in Example* 6.6.

Table 6.1

r	$\nu_2(\phi_{b_r})$	$\nu_2(\phi_{a^d_{t_r}})$	$\nu_2(\phi_{H_r})$	$\nu_2(\phi_{\mathcal{H}_r})$	$N(a^d_{t_r})$	$N(H_r)$	$N(\mathcal{H}_r)$
3	3.17513	1.17513	0.42927	1.17513	81	49	49
4	3.79313	1.79313	0.98084	1.79313	169	97	101
5	4.34408	2.34408	1.46708	2.34408	289	161	161
6	4.86202	2.86202	1.90387	2.86202	441	241	245
7	5.36283	3.36283	2.30033	3.36283	625	337	337
8	5.85293	3.85293	2.66264	3.85293	841	449	453
9	6.33524	4.33524	2.99578	4.33524	1089	577	577
10	6.81144	4.81144	3.30381	4.81144	1369	721	725
11	7.28260	5.28260	3.58991	5.28260	1681	881	881
12	7.74953	5.74953	3.85672	5.74953	2025	1057	1061

norm sense since $\nu_2(\phi_{\mathcal{H}_4}) \approx \nu_2(\phi_{b_4}) - \nu_1(\varphi_h)$. The graph and contour of $\phi_{\mathcal{H}_4}$ are presented in Figure 6.1.

Recall that by b_r we denote the interpolatory refinement mask supported on $[1-2r, 2r-1]$ as constructed by Deslauriers and Dubuc in [18]. By $\phi_{a^d_{t_r}}$ we denote the tensor product dual scaling function of φ_h with its mask $a^d_{t_r}$ satisfying

$$\overline{\widetilde{a_h}(z_1,z_2)}\widetilde{a^d_{t_r}}(z_1,z_2) = \overline{\widetilde{b_r}(z_1)}\widetilde{b_r}(z_2), \qquad (z_1,z_2) \in \mathbb{T}^2.$$

Let the masks H_r and $\mathcal{H}_r$ be the dual masks constructed by the TCBC algorithm and the modified TCBC algorithm, respectively, such that both H_r and $\mathcal{H}_r$ satisfy the sum rules of order $2r-2$. In the following, we use $N(a)$ to denote the number of nonzero coefficients in the refinement mask a. The values of $\nu_2(\phi_{b_r})$ are taken from [23]. Table 6.1 shows that for $r = 3, \ldots, 12$, the function $\phi_{\mathcal{H}_r}$ is an optimal dual scaling function of φ_h in the L_2 norm sense.

Acknowledgments. The author is indebted to Professor Rong-Qing Jia for his advice and comments on this paper and to Professor Zeev Ditzian for providing his reprints and sharing his ideas.

REFERENCES

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, *Image coding using wavelet transform*, IEEE Trans. Image Process, 12 (1992), pp. 205–220.

[2] J. Boman, *On a problem concerning moduli of smoothness*, in Proceedings of Fourier Analysis and Approximation, Colloq. Math. Soc. János Bolyai, Budapest, Hungary, 1976, pp. 175–179.

[3] C. de Boor, R. A. DeVore, and A. Ron, *On the construction of multivariate (pre)wavelets*, Constr Approx., 9 (1993), pp. 123–166.

[4] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 93, AMS, Providence, RI, 1991.

[5] C. K. Chui and C. Li, *A general framework of multivariate wavelets with duals*, Appl. Comput. Harmon. Anal., 1 (1994), pp. 368–390.

[6] C. K. Chui, J. Stöckler, and J. D. Ward, *Compactly supported box-spline wavelets*, Approx. Theory Appl., 8 (1992), pp. 77–100.

[7] C. K. Chui and J. Z. Wang, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–915.

[8] A. Cohen and I. Daubechies, *A stability criterion for biorthogonal wavelet bases and their related subband coding scheme*, Duke Math. J., 68 (1992), pp. 313–335.

[9] A. Cohen and I. Daubechies, *Nonseparable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9, (1993), pp. 51–137.

[10] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[11] A. Cohen, K. Gröchenig, and L. F. Villemoes, *Regularity of multivariate refinable functions*, Constr. Approx., 15 (1999), pp. 241–255.

[12] A. Cohen and J.-M. Schlenker, *Compactly supported bidimensional wavelets bases with hexagonal symmetry*, Constr. Approx., 9 (1993), pp. 209–236.

[13] D. Colella and C. Heil, *The characterization of continuous, four-coefficient scaling functions and wavelets*, IEEE Trans. Inform. Theory, 38 (1992), pp. 876–881.

[14] W. Dahmen and C. A. Micchelli, *Biorthogonal wavelet expansions*, Constr. Approx., 13 (1997), pp. 293–328.

[15] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 906–996.

[16] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[17] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[18] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.

[19] Z. Ditzian, *Moduli of continuity in $\mathbb{R}^n$ and $D \subset \mathbb{R}^n$*, Trans. Amer. Math. Soc., 282 (1984), pp. 611–623.

[20] Z. Ditzian, *Moduli of smoothness using discrete data*, J. Approx. Theory, 49 (1987), pp. 115–129.

[21] N. Dyn, *Subdivision schemes in computer-aided geometric design*, in Advances in Numerical Analysis II—Wavelets, Subdivision Algorithms and Radial Functions, W. A. Light, ed., Clarendon Press, Oxford, 1991, pp. 36–104.

[22] N. Dyn, J. A. Gregory, and D. Levin, *A butterfly subdivision scheme for surface interpolation with tension control*, ACM Trans. Graphics, 9 (1990), pp. 160–169.

[23] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[24] B. Han, *On dual wavelet tight frames*, Appl. Comput. Harmon. Anal., 4 (1997), pp. 380–413.

[25] B. Han and R. Q. Jia, *Multivariate refinement equations and convergence of subdivision schemes*, SIAM J. Math. Anal., 29 (1998), pp. 1177–1199.

[26] B. Han and R. Q. Jia, *Optimal interpolatory subdivision schemes in multidimensional spaces*, SIAM J. Numer. Anal., 36 (1999), pp. 105–124.

[27] H. Ji, S. D. Riemenschneider, and Z. Shen, *Multivariate compactly supported fundamental refinable functions, duals and biorthogonal wavelets*, Stud. Appl. Math., 102 (1999), pp. 173–204.

[28] R. Q. Jia, *Subdivision schemes in L_p spaces*, Adv. Comput. Math., 3 (1995), pp. 309–341.

[29] R. Q. Jia, *Characterization of smoothness of multivariate refinable functions in Sobolev spaces*, Trans. Amer. Math. Soc., 351 (1999), pp. 4089–4112.

[30] R. Q. Jia, *The subdivision and transition operators associated with a refinement equation*, in Advanced Topics in Multivariate Approximation, F. Fontanella, K. Jetter, and P.-J. Laurent, eds., World Scientific, River Edge, NJ, 1996, pp. 139–154.

[31] R. Q. Jia, *Approximation properties of multivariate wavelets*, Math. Comp., 67 (1998), pp. 647–665.

[32] R. Q. Jia, *Smoothness analysis of multivariate refinable functions and wavelets associated with arbitrary dilation matrices*, in preparation.

[33] R. Q. Jia and C. A. Micchelli, *Using the refinement equation for the construction of pre-wavelets* V: *Extensibility of trigonometric polynomials*, Computing, 48 (1992), pp. 61–72.

[34] R. Q. Jia and C. A. Micchelli, *On linear independence of integer translates of a finite number of functions*, Proc. Edinburgh Math. Soc. (2), 36 (1993), pp. 69–85.

[35] R. Q. Jia and Z. Shen, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc. (2), 37 (1994), pp. 271–300.

[36] J. Knipe, X. Li, and B. Han, *An improved lattice vector quantization scheme for wavelet compression*, IEEE Trans. Signal Process., 46 (1998), pp. 239–243.

[37] J. Kovačević and M. Vetterli, *Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$*, IEEE Trans. Inform. Theory, 38 (1992), pp. 533–555.

[38] W. Lawton, S. L. Lee, and Z. Shen, *An algorithm for matrix extension and wavelet construction*, Math. Comp., 65 (1996), pp. 723–737.

[39] P. G. Lemarié-Rieusset, *On the existence of compactly supported dual wavelets*, Appl. Comput. Harmon. Anal., 4 (1997), pp. 117–118.

[40] C. A. Micchelli, *Interpolatory subdivision schemes and wavelets*, J. Approx. Theory, 86 (1996), pp. 41–71.

[41] S. D. Riemenschneider and Z. Shen, *Multidimensional interpolatory subdivision schemes*, SIAM J. Numer. Anal., 34 (1997), pp. 2357–2381.

[42] S. D. Riemenschneider and Z. Shen, *Construction of Compactly Supported Biorthogonal Wavelets in* $L_2(\mathbb{R}^s)$, preprint, 1997.

[43] A. Ron and Z. Shen, *The Sobolev Regularity of Refinable Functions*, preprint, 1997.

[44] W. Sweldens, *The lifting scheme: A custom-design construction of biorthogonal wavelets*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 186–200.

[45] L. F. Villemoes, *Wavelet analysis of refinement equations*, SIAM J. Math. Anal., 25 (1994), pp. 1433–1460.

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 305–345

THE INFLUENCE OF LATERAL BOUNDARY CONDITIONS ON THE ASYMPTOTICS IN THIN ELASTIC PLATES*

MONIQUE DAUGE†, ISABELLE GRUAIS†, AND ANDREAS RÖSSLE‡

Abstract. Here we investigate the limits and the boundary layers of the three-dimensional displacement in thin elastic plates as the thickness tends to zero in each of the eight main types of lateral boundary conditions on their edges: hard and soft clamped, hard and soft simple support, friction conditions, sliding edge, and free plates. Relying on construction algorithms [M. Dauge and I. Gruais, *Asymptotic Anal.*, 13 (1996), pp. 167–197], we establish an asymptotics of the displacement combining inner and outer expansions. We describe the two first terms in the outer expansion: these are Kirchhoff–Love displacements satisfying prescribed boundary conditions that we exhibit. We also study the first boundary layer term: when the transverse component is clamped, it has generically nonzero transverse and normal components, whereas when the transverse component is free, the first boundary layer term is of bending type and has only its nonzero in-plane tangential component.

Key words. thin plates, linear elasticity, singular perturbation, boundary layer, asymptotic expansion

AMS subject classifications. 73K10, 73C35, 35J25, 35B25

PII. S0036141098333025

1. Introduction.

The problem of thin elastic plate bending in linearized elastostatics has been addressed for more than 150 years (the first correct model was presented in a paper by Kirchhoff [18] published in 1850). But due to the singular perturbation nature of the problem as the thickness of the plate tends to zero, it is not quite straightforward to perform a rigorous mathematical analysis of characteristic fields and tensors, solutions of the three-dimensional equations. However, the knowledge of accurate asymptotics allows first an evaluation of the validity of mechanical models, and second, the construction of simplified and performing numerical models.

In the case when the plate is clamped along its lateral boundary, the situation is now well-known, at least theoretically: The comparison between three-dimensional and two-dimensional models was first performed by the construction of infinite *formal* asymptotic expansions; see Friedrichs and Dressler [15], Gol'denveizer [16], Gregory and Wan [17]. Shortly before, Morgenstern [21] was indeed the first to prove that the Kirchhoff model [18] is the correct asymptotic limit of the three-dimensional model when the thickness approaches zero in the hard clamped, hard simply supported and free plate situations by using the Prager–Synge hypercircle theorem [29]. Next, rigorous error estimates between the three-dimensional solution and its limit were proved by Shoikhet [31] and by Ciarlet and Destuynder [5, 13, 3]. Further terms were exhibited by Nazarov and Zorin [24], and the whole asymptotic expansion was constructed in [8, 9].

*Received by the editors January 27, 1998; accepted for publication (in revised form) December 10, 1998; published electronically December 21, 1999.
http://www.siam.org/journals/sima/31-2/33302.html

†Institut Mathématique, UMR 6625 du CNRS, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes, France (dauge@univ-rennes1.fr, gruais@univ-rennes1.fr).

‡Mathematical Institute A/6, University of Stuttgart, Pfaffenwaldring 57, 70550 Stuttgart, Germany (roessle@mathematik.uni-stuttgart.de). The research of this author was supported by a grant of the German National Science Foundation (DFG, graduate collegium "GKKS – Modelling and discretization methods in continua and fluids").

Different types of lateral boundary conditions are of interest: let us quote the soft clamped plate where the tangential in-plane component of the displacement is free, the hard simply supported plate where its normal component is free, the soft simply supported plate where both above components of the displacement are free, and also the totally free plate. These cases were investigated by Arnold and Falk [1], where an asymptotics for the Reissner–Mindlin plate was constructed, and by Chen [2], where error bounds between the three-dimensional solution and its limit were evaluated.

In this paper, we prove the validity of an infinite asymptotic expansion of the displacement with optimal error estimates in H^1, L^2, and energy norms. Such an expansion can be differentiated and provides corresponding results for the stress and the strain tensors; see [7] for the clamped case. As in [24] and [8, 9], this asymptotics includes, simultaneously,

- an outer part containing displacements only depending on the in-plane variables x_* and on the scaled transverse variable x_3;
- an inner part containing exponentially decaying profiles (boundary layer terms), depending on two scaled variables (x_3 and $t = r/\varepsilon$, where r is the distance to the lateral boundary).

As material law, we choose to remain in the framework of homogeneous, isotropic materials, which allows us to uncouple the boundary layer terms $\boldsymbol{\varphi}$ into two parts:

- the horizontal tangential component φ_s governed by the Laplace equation,
- the two other components φ_t and φ_3 governed by the bidimensional Lamé equations, whose solutions can themselves be uncoupled in membrane and bending modes, i.e., possessing parity properties with respect to the transverse variable: the former having an even φ_t and an odd φ_3 and the latter having converse properties.

Thus, conditions ensuring the exponential decay at infinity of solutions of the above problems can be made explicit, resulting in simple coupling formulas between the inner and outer parts of the expansion. These coupling formulas lead to the determination of boundary conditions for the limit membrane and bending problems.

The first boundary layer terms bring the quantitative limitation of accuracy of bidimensional models. In the clamped and simple support cases, we find a strong boundary layer term with generically nonzero membrane and bending parts, whereas in the frictional and free cases, we find a first boundary layer term which has the bending type and only the in-plane tangential component nonzero, and moreover, the subprincipal term in the outer part of the expansion is a Kirchhoff–Love displacement as usual, but with zero membrane part. Thus if the right-hand side has the membrane type, the solution of the three-dimensional Lamé equations for the free plate converges to the usual limit Kirchhoff–Love displacement with improved accuracy.

This paper contains twelve sections: in section 1 we introduce the elasticity problems and in section 2 we present our results in the form of several tables. In section 3 we give as an algorithm the construction rules for the outer part of the Ansatz, while in section 4 we formulate the boundary value problems on a half-strip governing the boundary layer profiles $\boldsymbol{\varphi}$, and in section 5 we give the conditions on the data ensuring the existence of exponentially decreasing solutions to these problems. The five next sections are devoted to each of the eight types of lateral boundary conditions with more emphasis on five of them: hard and soft clamped, hard simple support, sliding edge and free plates. In section 11, we prove error estimates between the three-dimensional solution and any truncated series from the asymptotic expansion, and analyze the regularity of the different terms in the asymptotics: whereas the outer expansion terms are smooth if the data are so, the profiles have singularities along the edges of the plate. We conclude in section 12 with considerations about relative errors between the three-dimensional solution and a limited two-dimensional solution, which has to be carefully chosen according to what we wish to approximate (the displacement in H^1 norm, or the strain in L^2 norm).

2. Lateral boundary conditions. We aim to study the behavior of the displacement field $\boldsymbol{u}^\varepsilon$ in a family of thin elastic three-dimensional plates Ω^ε as the thickness parameter ε tends to zero. The plates Ω^ε are constituted of a homogeneous, isotropic material with Lamé constants and μ, and are defined as follows:

$$\Omega^\varepsilon = \omega \times (-\varepsilon, +\varepsilon) \quad \text{with } \ \omega \subset \mathbb{R}^2 \ \text{ a regular domain and } \varepsilon > 0\,.$$

Let $\Gamma^\varepsilon_\pm$ be their upper and lower faces $\omega \times \{\pm\varepsilon\}$ and Γ^ε_0 be their lateral faces $\partial\omega \times (-\varepsilon, +\varepsilon)$.

2.1. Cartesian, scaled, and local coordinates. Let $\tilde{x} = (x_1,\, x_2,\, \tilde{x}_3)$ be the Cartesian coordinates in the plates Ω^ε. We will often denote by x_* the in-plane coordinates $(x_1,\, x_2) \in \omega$ and by α or β the indices in $\{1, 2\}$ corresponding to the in-plane variables. The dilatation along the vertical axis ($x_3 = \varepsilon^{-1}\,\tilde{x}_3$) transforms Ω^ε into the fixed reference configuration $\Omega = \omega \times (-1, +1)$:

$$\tilde{x} = (x_*,\, \tilde{x}_3) \in \Omega^\varepsilon = \omega \times (-\varepsilon, +\varepsilon) \longmapsto x = (x_*,\, x_3) \in \Omega = \omega \times (-1, +1). \tag{2.1}$$

We also have to introduce in-plane local coordinates (r, s) in a neighborhood of the boundary $\partial\omega$. Let $\boldsymbol{n}$ be the inner unit normal to $\partial\omega$ and $\boldsymbol{\tau}$ be the tangent unit vector field to $\partial\omega$ such that the basis $(\boldsymbol{\tau}, \boldsymbol{n})$ is direct in each point of $\partial\omega$. Denote by s a curvilinear abscissa (arc length) along $\partial\omega$ oriented according to $\boldsymbol{\tau}$. Let $\mathbb{S} \sim \partial\omega$ be the set of the values of s:

$$\mathbb{S} \ni s \longmapsto \boldsymbol{\gamma}(s) \in \partial\omega.$$

For a point $x_* \in \mathbb{R}^2$, let $r = r(x_*)$ be its signed distance to $\partial\omega$ oriented along $\boldsymbol{n}$; i.e., r is this distance if $x_* \in \omega$ and r is minus this distance if $x_* \notin \omega$. If $|r|$ is small enough, there exists a unique point $x^0_* \in \partial\omega$ such that $|r| = \operatorname{dist}(x_*, x^0_*)$ and we define $s = s(x_*)$ as the curvilinear abscissa of x^0_*. Thus, we have a tubular neighborhood of $\partial\omega$ which is diffeomorphic to $(-r^0, r^0) \times \mathbb{S}$ via the change of variables $x_* \mapsto (r, s)$. And, in this tubular neighborhood, the partial derivatives ∂_r and ∂_s are well defined (and, of course, commute with each other).

We extend the vector fields $\boldsymbol{n}$ and $\boldsymbol{\tau}$ from $\mathbb{S}$ to $(-r^0, r^0) \times \mathbb{S}$ by

$$\forall r \in (-r^0, r^0), \quad \forall s \in \mathbb{S}, \qquad \boldsymbol{n}(r, s) = \boldsymbol{n}(s) \quad \text{and} \quad \boldsymbol{\tau}(r, s) = \boldsymbol{\tau}(s).$$

We have

$$\boldsymbol{n} = \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\tau} = \begin{pmatrix} n_2 \\ -n_1 \end{pmatrix}.$$

Moreover, with $R = R(s)$ the curvature radius of $\partial\omega$ at s from inside ω and $\kappa = \frac{1}{R}$ the curvature, there holds (the last identities are Frenet's relations)

$$\partial_r \boldsymbol{n} = 0, \quad \partial_r \boldsymbol{\tau} = 0 \quad \text{and} \quad \partial_s \boldsymbol{n} = -\,\kappa\,\boldsymbol{\tau}, \quad \partial_s \boldsymbol{\tau} = \,\kappa\,\boldsymbol{n}.$$

Thus, relying on the relation $x_* = \boldsymbol{\gamma}(s) + r\,\boldsymbol{n}(s)$, we obtain

$$\partial_r = n_1\partial_1 + n_2\partial_2 \quad \text{and} \quad \partial_s = (1 - \kappa\, r)(n_2\partial_1 - n_1\partial_2). \tag{2.2}$$

Of course $\partial_n = \partial_r$.

Table 1
Lateral boundary conditions.

ⓘ	Type	Dirichlet	Neumann	$A_{\textcircled{\scriptsize i}}$	$B_{\textcircled{\scriptsize i}}$
①	hard clamped	$\boldsymbol{u}=0,$		$\{n,s,3\}$	
②	soft clamped	$u_n,\ u_3=0,$	$T_s=0$	$\{n,3\}$	$\{s\}$
③	hard simply supported	$u_s,\ u_3=0,$	$T_n=0$	$\{s,3\}$	$\{n\}$
④	soft simply supported	$u_3=0,$	$T_n,\ T_s=0$	$\{3\}$	$\{n,s\}$
⑤	frictional I	$u_n,\ u_s=0,$	$T_3=0$	$\{n,s\}$	$\{3\}$
⑥	sliding edge	$u_n=0,$	$T_s,\ T_3=0$	$\{n\}$	$\{s,3\}$
⑦	frictional II	$u_s=0,$	$T_n,\ T_3=0$	$\{s\}$	$\{n,3\}$
⑧	free		$\boldsymbol{T}=0$		$\{n,s,3\}$

2.2. Cartesian, scaled, and local tensors. The displacement and traction tensors in Ω^ε are denoted $\boldsymbol{u}^\varepsilon$ and $\boldsymbol{T}^\varepsilon$ and their Cartesian components are $(u_1^\varepsilon, u_2^\varepsilon, u_3^\varepsilon)$ and $(T_1^\varepsilon, T_2^\varepsilon, T_3^\varepsilon)$. As $\boldsymbol{u}$ is covariant, it is naturally transformed by the scaling (2.1) into $\boldsymbol{u}(\varepsilon)$ according to

$$u_\alpha(\varepsilon)(x) = u_\alpha^\varepsilon(\tilde{x}),\ \alpha=1,2, \qquad u_3(\varepsilon)(x) = \varepsilon\, u_3^\varepsilon(\tilde{x}), \tag{2.3}$$

whereas $\boldsymbol{T}$, which is contravariant, is transformed according the same laws as the volume force field $\boldsymbol{f}^\varepsilon$: by the scaling (2.1) $\boldsymbol{f}^\varepsilon$ is transformed into $\boldsymbol{f}(\varepsilon)$

$$f_\alpha(\varepsilon)(x) = f_\alpha^\varepsilon(\tilde{x}),\ \alpha=1,2, \qquad f_3(\varepsilon)(x) = \varepsilon^{-1} f_3^\varepsilon(\tilde{x}). \tag{2.4}$$

In the tubular neighborhood $(-r^0, r^0)\times\mathbb{S}$, in view of (2.2) we can introduce the in-plane normal and tangential components of $\boldsymbol{u}$ and $\boldsymbol{T}$ by

$$u_n = n_1u_1 + n_2u_2 \quad\text{and}\quad u_s = (1-\kappa\, r)(n_2u_1 - n_1u_2), \tag{2.5a}$$

$$T_n = n_1T_1 + n_2T_2 \quad\text{and}\quad T_s = (1-\kappa\, r)^{-1}(n_2T_1 - n_1T_2). \tag{2.5b}$$

2.3. The equations of elasticity. As standard, let $e(\boldsymbol{u})$ denote the linearized strain tensor $e_{ij}(\boldsymbol{u}) = \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right)$ associated with the displacement $\boldsymbol{u}$. Then the stress tensor $\sigma(\boldsymbol{u})$ is given by Hooke's law $\sigma(\boldsymbol{u}) = A\, e(\boldsymbol{u})$, where the rigidity matrix $A = (A_{ijkl})$ of the material is given by $A_{ijkl} = \lambda\,\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$. The inward traction field at a point on the boundary is $\boldsymbol{T}$ defined as $\sigma(\boldsymbol{u})\,\boldsymbol{n}$, where $\boldsymbol{n}$ is the unit interior normal to the boundary.

We make the assumption that the boundary conditions on the upper and lower faces $\Gamma_\pm^\varepsilon$ of the plate are of traction type. On the lateral face Γ_0^ε we are going to consider the eight "canonical" choices of boundary conditions which will be denoted by ⓘ, where $\mathbf{i} = 1,\dots,8$. Indeed, on the lateral boundary Γ_0^ε we can distinguish three natural components in the displacements or the tractions: normal, horizontal tangential, and vertical, and we obtain eight "canonical" lateral boundary conditions, according to how we choose to prescribe the displacement or the traction for each component.

On Γ_0^ε, we recall that the normal component of $\boldsymbol{u}$ is $u_n = u_1n_1 + u_2n_2$, its horizontal tangential component is $u_s = u_1n_2 - u_2n_1$, and its vertical component is u_3. Similar notations apply to $\boldsymbol{T}$. Each boundary condition ⓘ corresponds to two complementary sets of indices $A_{\textcircled{\scriptsize i}}$ and $B_{\textcircled{\scriptsize i}}$, where $A_{\textcircled{\scriptsize i}}$ is attached to the Dirichlet conditions of $\textcircled{i}$; i.e., $u_a = 0$ for each index $a \in A_{\textcircled{i}}$: these are the *stable* conditions. The Neumann conditions are $T_b = 0$ for each index $b \in B_{\textcircled{i}}$ and appear as *natural* conditions.

To each boundary condition $\textcircled{i}$ is associated the space of displacements $V_{\textcircled{i}}(\Omega^\varepsilon)$ of the $\boldsymbol{v} \in H^1(\Omega^\varepsilon)^3$ such that $v_a = 0$ for all $a \in A_{\textcircled{i}}$, and the space $\mathcal{R}_{\textcircled{i}}$ of the rigid motions satisfying the Dirichlet conditions of $V_{\textcircled{i}}$. Then, the variational formulation of the problem consists of finding

$$(2.6)\quad \begin{cases} \boldsymbol{u}^\varepsilon \in V_{\textcircled{i}}(\Omega^\varepsilon) \\ \forall\, \boldsymbol{v} \in V_{\textcircled{i}}(\Omega^\varepsilon), \quad \displaystyle\int_{\Omega^\varepsilon} A\, e(\boldsymbol{u}^\varepsilon) : e(\boldsymbol{v}) = \int_{\Omega^\varepsilon} \boldsymbol{f}^\varepsilon \cdot \boldsymbol{v} \, + \int_{\Gamma_+^\varepsilon} \boldsymbol{g}^{\varepsilon,+} \cdot \boldsymbol{v} \, - \int_{\Gamma_-^\varepsilon} \boldsymbol{g}^{\varepsilon,-} \cdot \boldsymbol{v}, \end{cases}$$

where $\boldsymbol{f}^\varepsilon$ represents the volume force and $\boldsymbol{g}^{\varepsilon,\pm}$ the prescribed horizontal tractions. If the right-hand side satisfies the correct compatibility condition (orthogonality to all $\boldsymbol{v} \in \mathcal{R}_{\textcircled{i}}(\Omega^\varepsilon)$), then there exists a unique solution to (2.6) satisfying the orthogonality conditions $\int_{\Omega^\varepsilon} \boldsymbol{u}^\varepsilon \cdot \boldsymbol{v} = 0$ for all $\boldsymbol{v} \in \mathcal{R}_{\textcircled{i}}(\Omega^\varepsilon)$.

After the scaling (2.3), an asymptotic expansion of $\boldsymbol{u}(\varepsilon)$ makes sense if the scaled data have comparable behaviors as ε is varying. To this aim, we make the assumption on the right-hand sides that they are given by profiles in x_3, namely,

$$(2.7\text{a})\qquad f^\varepsilon_\alpha(\tilde{x}) = f_\alpha(x),\ \alpha = 1,2, \qquad \varepsilon^{-1} f^\varepsilon_3(\tilde{x}) = f_3(x),$$

$$(2.7\text{b})\qquad \varepsilon^{-1} g^{\varepsilon,\pm}_\alpha(\tilde{x}) = g^{\pm}_\alpha(x_*),\ \alpha = 1,2, \qquad \varepsilon^{-2} g^{\varepsilon,\pm}_3(\tilde{x}) = g^{\pm}_3(x_*);$$

compare with (2.4) for the homogeneities. To simplify, we assume that the profiles $\boldsymbol{f}$ and $\boldsymbol{g}^\pm$ are regular up to the boundary, i.e., $\boldsymbol{f} \in \mathcal{C}^\infty(\overline{\Omega})^3$ and $\boldsymbol{g}^\pm \in \mathcal{C}^\infty(\overline{\omega})^3$.

After scaling (2.3) and assumption (2.7), problem (2.6) is transformed into a new boundary value problem on Ω, where now the operators depend on the small parameter ε: The variational formulation of the problem for the scaled displacement $\boldsymbol{u}(\varepsilon)$ consists of finding

$$(2.8)\quad \begin{cases} \boldsymbol{u}(\varepsilon) \in V_{\textcircled{i}}(\Omega), \\ \forall\, \boldsymbol{v} \in V_{\textcircled{i}}(\Omega), \quad \displaystyle\int_{\Omega} A\, \theta(\varepsilon)(\boldsymbol{u}(\varepsilon)) : \theta(\varepsilon)(\boldsymbol{v}) = \int_{\Omega} \boldsymbol{f} \cdot \boldsymbol{v} \, + \int_{\Gamma_+} \boldsymbol{g}^{+} \cdot \boldsymbol{v} \, - \int_{\Gamma_-} \boldsymbol{g}^{-} \cdot \boldsymbol{v}, \end{cases}$$

where $V_{\textcircled{i}}(\Omega)$ is the space of the geometrically admissible displacements $\boldsymbol{v} \in H^1(\Omega)^3$ associated with the problem with lateral boundary conditions $\textcircled{i}$, and $\theta(\varepsilon)(\boldsymbol{v})$ denotes the scaled linearized strain tensor defined by

$$(2.9)\quad \theta_{\alpha\beta}(\varepsilon)(\boldsymbol{v}) := e_{\alpha\beta}(\boldsymbol{v})\,, \quad \theta_{\alpha 3}(\varepsilon)(\boldsymbol{v}) := \varepsilon^{-1}\, e_{\alpha 3}(\boldsymbol{v})\,, \quad \theta_{33}(\varepsilon)(\boldsymbol{v}) := \varepsilon^{-2}\, e_{33}(\boldsymbol{v})$$

for $\alpha, \beta = 1, 2$; note that there holds $\theta(\varepsilon)(\boldsymbol{u}(\varepsilon)) = e(\boldsymbol{u}^\varepsilon)$.

Denoting by $\mathcal{R}_{\textcircled{i}}(\Omega)$ the space of rigid motions satisfying the Dirichlet conditions of $V_{\textcircled{i}}(\Omega)$, the compatibility condition becomes

$$(2.10)\qquad \forall\, \boldsymbol{v} \in \mathcal{R}_{\textcircled{i}}(\Omega), \quad \int_{\Omega} \boldsymbol{f} \cdot \boldsymbol{v} \, + \int_{\Gamma_+} \boldsymbol{g}^{+} \cdot \boldsymbol{v} \, - \int_{\Gamma_-} \boldsymbol{g}^{-} \cdot \boldsymbol{v} = 0$$

and $\boldsymbol{u}(\varepsilon)$ satisfies the orthogonality condition

$$(2.11)\qquad \forall\, \boldsymbol{v} \in \mathcal{R}_{\textcircled{i}}(\Omega), \quad \int_{\Omega} \boldsymbol{u}(\varepsilon)\, \cdot\, \boldsymbol{v} = 0\,.$$

Problem (2.8) can be written in the boundary value problem form (2.12a)–(2.14) on Ω as follows. To formulate it, we use the repeated index convention. Moreover, $\boldsymbol{u}_*$ is a condensed notation for (u_1, u_2), $\operatorname{div}_* \boldsymbol{u}_*$ denotes $\partial_1 u_1 + \partial_2 u_2$, and Δ_* denotes the horizontal Laplacian $\partial_{11} + \partial_{22}$. The in-plane components, indexed by $\alpha = 1, 2$, and the vertical component of the interior equations in Ω are

$$2\mu\, \partial_3 e_{\alpha 3}(\boldsymbol{u}) + \lambda\, \partial_{\alpha 3} u_3 + \varepsilon^2 \big((\lambda + \mu)\partial_\alpha \operatorname{div}_* \boldsymbol{u}_* + \mu\, \Delta_* u_\alpha\big) = -\varepsilon^2 f_\alpha, \tag{2.12a}$$

$$(\lambda + 2\mu)\partial_{33} u_3 + \varepsilon^2 \big(\lambda\, \partial_3 \operatorname{div}_* \boldsymbol{u}_* + 2\mu\, \partial_\beta e_{\beta 3}(\boldsymbol{u})\big) = -\varepsilon^4 f_3. \tag{2.12b}$$

The boundary conditions on the horizontal sides $\Gamma_\pm := \{x_3 = \pm 1\} \cap \partial\Omega$ are

$$2\mu\, e_{\alpha 3}(\boldsymbol{u}) = \varepsilon^2 g_\alpha^{\pm}\,, \qquad \alpha = 1, 2, \tag{2.13a}$$

$$(\lambda + 2\mu)\partial_3 u_3 + \varepsilon^2\, \lambda \operatorname{div}_* \boldsymbol{u}_* = \varepsilon^4 g_3^{\pm}\,. \tag{2.13b}$$

The boundary conditions on the lateral side $\Gamma_0 = \partial\omega \times (-1, 1)$ can be written as

$$u_a = 0, \quad \forall a \in A_{\textcircled{i}} \qquad \text{and} \qquad T_b = 0, \quad \forall b \in B_{\textcircled{i}}. \tag{2.14}$$

The normal, tangential horizontal, and vertical components of the traction $\boldsymbol{T} = \boldsymbol{T}(\varepsilon)$ on Γ_0 are given by, respectively,

$$T_n(\varepsilon) = \lambda\, \partial_3 u_3(\varepsilon) + \varepsilon^2 \big(\lambda \operatorname{div}_* \boldsymbol{u}_*(\varepsilon) + 2\mu\, \partial_n u_n(\varepsilon)\big), \tag{2.15a}$$

$$T_s(\varepsilon) = \varepsilon^2 \mu\big(\partial_s u_n(\varepsilon) + \partial_n u_s(\varepsilon) + 2\kappa\, u_s(\varepsilon)\big), \tag{2.15b}$$

$$T_3(\varepsilon) = \mu\big(\partial_n u_3(\varepsilon) + \partial_3 u_n(\varepsilon)\big). \tag{2.15c}$$

3. Description of results.

We first state the common features of the asymptotics of the scaled displacement $\boldsymbol{u}(\varepsilon)$ and next deduce the asymptotics of the displacement $\boldsymbol{u}^\varepsilon$ in the thin plates. Then we describe the first terms of the asymptotics in each of the eight lateral boundary conditions.

3.1. Common features.

Just as in the well-known situation of the clamped plate, the scaled displacement $\boldsymbol{u}(\varepsilon)$ tends in Ω to a Kirchhoff–Love displacement. Let us recall the following definition.

Definition 3.1. *A displacement $\boldsymbol{u}$ in Ω is called a* Kirchhoff–Love displacement *if there exist a displacement $\boldsymbol{\zeta}_* = (\zeta_1, \zeta_2)$ in the mean surface ω and a function ζ_3 on ω such that*

$$\boldsymbol{u} = (\zeta_1 - x_3 \partial_1 \zeta_3,\ \zeta_2 - x_3 \partial_2 \zeta_3,\ \zeta_3).$$

The function $\boldsymbol{\zeta} := (\boldsymbol{\zeta}_*, \zeta_3)$ is called the generator *of $\boldsymbol{u}$, and the descaled displacement associated with $\boldsymbol{u}$ in Ω^ε has exactly the same form with x_3 replaced by $\tilde{x}_3$. Then*

$$\boldsymbol{u}_{\mathrm{KL,b}} = (-x_3 \partial_1 \zeta_3, -x_3 \partial_2 \zeta_3,\ \zeta_3) \quad \textit{and} \quad \boldsymbol{u}_{\mathrm{KL,m}} = (\zeta_1, \zeta_2, 0) \tag{3.1}$$

are respectively the bending and membrane parts of $\boldsymbol{u}$.

The asymptotics of $\boldsymbol{u}(\varepsilon)$ contains three types of terms for $k \geq 0$:

- $\boldsymbol{u}_{\mathrm{KL}}^k$: Kirchhoff–Love displacements with 'generating functions' $\boldsymbol{\zeta}^k = (\boldsymbol{\zeta}_*^k, \zeta_3^k)$, i.e., $\boldsymbol{u}_{\mathrm{KL}}^k(x) = \big(\boldsymbol{\zeta}_*^k(x_*) - x_3 \nabla_* \zeta_3^k(x_*), \zeta_3^k(x_*)\big)$,
- $\boldsymbol{v}^k$: displacements with zero mean value: $\forall x_* \in \overline{\omega}$, $\int_{-1}^{+1} \boldsymbol{v}^k(x_*, x_3)\, dx_3 = 0$,
- $\boldsymbol{w}^k$: exponentially decreasing profiles as $t \to +\infty$

and can be written as

$$\boldsymbol{u}(\varepsilon)(x) \simeq \boldsymbol{u}^0_{\mathrm{KL}} + \varepsilon \boldsymbol{u}^1\left(x, \frac{r}{\varepsilon}\right) + \cdots + \varepsilon^k \boldsymbol{u}^k\left(x, \frac{r}{\varepsilon}\right) + \cdots, \tag{3.2}$$

where

$$\begin{array}{rclll} \boldsymbol{u}^1(x,t) & = & \boldsymbol{u}^1_{\mathrm{KL}} \phantom{+ \boldsymbol{v}^k} + \chi(r)\,\boldsymbol{w}^1(t,s,x_3) & \text{with} & w^1_3 = 0, \\ \boldsymbol{u}^k(x,t) & = & \boldsymbol{u}^k_{\mathrm{KL}} + \boldsymbol{v}^k + \chi(r)\,\boldsymbol{w}^k(t,s,x_3) & \text{for} & k \geq 2\,, \end{array} \tag{3.3}$$

with χ a cut-off function equal to 1 in a neighborhood of $\partial\omega$.

Theorem 3.2. *Let $\boldsymbol{u}(\varepsilon)$ be the unique solution of problem* (2.8) *satisfying the mean value conditions* (2.11). *Then there exist Kirchhoff–Love generators $\boldsymbol{\zeta}^k$ for $k \geq 0$, displacements with zero mean value $\boldsymbol{v}^k$ for $k \geq 2$, and profiles $\boldsymbol{w}^k$ for $k \geq 1$ such that there holds $\forall N \geq 0$*

$$\left\| \boldsymbol{u}(\varepsilon)(x) - \boldsymbol{u}^0_{\mathrm{KL}}(x) - \sum_{k=1}^{N} \varepsilon^k \boldsymbol{u}^k\ \left(x, \frac{r}{\varepsilon}\right) \right\| H^1(\Omega)^3 \leq C\,\varepsilon^{N+1/2} \tag{3.4}$$

with $\boldsymbol{u}^k(x, \frac{r}{\varepsilon})$ given in (3.3).

Let us point out that the "physical" displacement $\boldsymbol{u}^\varepsilon$ expands like $\boldsymbol{u}(\varepsilon)$ in the following way in the sense of asymptotic expansions:

$$\begin{aligned} \boldsymbol{u}^\varepsilon \simeq \boldsymbol{u}^0_{\mathrm{KL,b}} + \tilde{\boldsymbol{u}}^0_{\mathrm{KL,m}} + \tilde{\boldsymbol{u}}^0_{\mathrm{KL,b}} + \varepsilon(\tilde{\boldsymbol{u}}^1_{\mathrm{KL,m}} + \tilde{\boldsymbol{u}}^1_{\mathrm{KL,b}} + \tilde{\boldsymbol{v}}^1 + \boldsymbol{\varphi}^1) + \cdots \\ + \varepsilon^k(\tilde{\boldsymbol{u}}^k_{\mathrm{KL,m}} + \tilde{\boldsymbol{u}}^k_{\mathrm{KL,b}} + \tilde{\boldsymbol{v}}^k + \boldsymbol{\varphi}^k) + \cdots, \end{aligned} \tag{3.5}$$

where

- $\tilde{\boldsymbol{u}}^k_{\mathrm{KL,b}}$ and $\tilde{\boldsymbol{u}}^k_{\mathrm{KL,m}}$ are the bending and membrane parts on Ω^ε of the Kirchhoff–Love displacement with generator $\boldsymbol{\zeta}^k$;
- $\tilde{\boldsymbol{v}}^k = \tilde{\boldsymbol{v}}^k(x_*, \frac{\tilde{x}_3}{\varepsilon})$, i.e., does not depend on ε in the scaled domain Ω;
- $\boldsymbol{\varphi}^k = \boldsymbol{\varphi}^k(\frac{r}{\varepsilon}, s, \frac{\tilde{x}_3}{\varepsilon})$ is a boundary layer profile.

The links with expansion (3.2) on the thin plates are simply provided by the following relations:

$$\left\{ \begin{array}{l} \tilde{\boldsymbol{u}}^k_{\mathrm{KL,b}}(\tilde{x}) = \boldsymbol{u}^k_{\mathrm{KL,b}}(x), \quad \tilde{\boldsymbol{u}}^k_{\mathrm{KL,m}}(\tilde{x}) = \boldsymbol{u}^k_{\mathrm{KL,m}}(x), \\ \tilde{\boldsymbol{v}}^k = (\boldsymbol{v}^k_*, v^{k+1}_3), \quad \text{and} \quad \boldsymbol{\varphi}^k = (\boldsymbol{w}^k_*, w^{k+1}_3)\,. \end{array} \right. \tag{3.6}$$

In Table 8, we will give the formulas linking the displacements $\mathbf{v}^k$ to the Kirchhoff–Love generators. These formulas do not depend on the nature of the lateral boundary conditions. In particular, the first non-Kirchhoff displacement $\tilde{\boldsymbol{v}}^1 = (\mathbf{0}, v^2_3)$ is completely determined by $\boldsymbol{\zeta}^0$, *cf* Destuynder [14] for a similar formula:

$$\tilde{\boldsymbol{v}}^1(x_*, x_3) = \frac{\lambda}{6(\lambda + 2\mu)}\Big(0,\ 0,\ -6x_3 \operatorname{div}_* \boldsymbol{\zeta}^0_* + (3x_3^2 - 1)\,\Delta_* \zeta^0_3\Big). \tag{3.7}$$

3.2. Specific features: The Kirchhoff–Love generators. The generators $\boldsymbol{\zeta}^k_*$ and ζ^k_3 of the above Kirchhoff displacements are solutions of membrane and bending equations, respectively, with boundary conditions on $\partial\omega$. Let us first write down these membrane and bending operators together with the associated Dirichlet and Neumann conditions. Then we describe the boundary operators and data associated with the generators.

3.2.1. Membrane. The bilinear form associated with the membrane operator L^{m} (plane stress model)

$$L^{\mathrm{m}}\boldsymbol{\zeta}_* = \mu\,\boldsymbol{\Delta}_*\boldsymbol{\zeta}_* + (\tilde\lambda+\mu)\nabla_*\operatorname{div}_*\boldsymbol{\zeta}_* \tag{3.8}$$

is $\int_\omega \tilde\lambda\, e_{\alpha\alpha}(\boldsymbol{\zeta}_*)\; e_{\beta\beta}(\boldsymbol{\eta}_*) + 2\mu\, e_{\alpha\beta}(\boldsymbol{\zeta}_*)\; e_{\alpha\beta}(\boldsymbol{\eta}_*)$ with the homogenized Lamé coefficient

$$\tilde\lambda = \frac{2\lambda\mu}{\lambda+2\mu}\,. \tag{3.9}$$

In normal and tangential components (cf. (2.5a))

$$\zeta_n = n_1\zeta_1 + n_2\zeta_2 \quad\text{and}\quad \zeta_s = (1-\kappa\, r)(n_2\zeta_1 - n_1\zeta_2),$$

the Dirichlet traces are simply (ζ_n, ζ_s) on $\partial\omega$, and the Neumann traces are

$$T^{\mathrm{m}}_n(\boldsymbol{\zeta}_*) = \tilde\lambda\operatorname{div}_*\boldsymbol{\zeta}_* + 2\mu\,\partial_n\zeta_n, \tag{3.10a}$$

$$T^{\mathrm{m}}_s(\boldsymbol{\zeta}_*) = \mu(\partial_s\zeta_n + \partial_n\zeta_s + 2\kappa\,\zeta_s), \tag{3.10b}$$

where ∂_n and ∂_s are defined in (2.2).

3.2.2. Bending. The bilinear form associated with the bending operator L^{b},

$$L^{\mathrm{b}}\zeta_3 = (\tilde\lambda+2\mu)\Delta^2_*\zeta_3 \tag{3.11}$$

is $\int_\omega \tilde\lambda\,\partial_{\alpha\alpha}\zeta_3\;\partial_{\beta\beta}\eta_3 + 2\mu\,\partial_{\alpha\beta}\zeta_3\;\partial_{\alpha\beta}\eta_3$. Its Dirichlet traces are ζ_3 and $\partial_n\zeta_3$ on $\partial\omega$, whereas the Neumann traces are

$$M_n(\zeta_3) = \tilde\lambda\,\Delta_*\zeta_3 + 2\mu\,\partial_{nn}\zeta_3, \tag{3.12a}$$

$$N_n(\zeta_3) = (\tilde\lambda+2\mu)\partial_n\Delta_*\zeta_3 + 2\mu\,\partial_s(\partial_n+\kappa)\partial_s\zeta_3. \tag{3.12b}$$

The mechanical interpretation of these boundary operators is that M_n corresponds to the “Kirchhoff bending moment” and N_n corresponds to the “Kirchhoff shear force” on the lateral side of the plate (up to constants only depending on λ and μ).

3.2.3. Boundary value problems for the Kirchhoff–Love generators. The $\boldsymbol{\zeta}^k_*$ and ζ^k_3 are solutions of equations of the type

$$L^{\mathrm{m}}(\boldsymbol{\zeta}^k_*) = \boldsymbol{R}^k_{\mathrm{m}} \quad\text{in } \omega, \qquad \gamma^{\mathrm{m},1}(\boldsymbol{\zeta}^k_*) = \gamma^k_{\mathrm{m},1}, \;\text{ and }\; \gamma^{\mathrm{m},2}(\boldsymbol{\zeta}^k_*) = \gamma^k_{\mathrm{m},2} \quad\text{on } \partial\omega, \tag{3.13a}$$

$$L^{\mathrm{b}}(\zeta^k_3) = R^k_{\mathrm{b}} \quad\text{in } \omega, \qquad \gamma^{\mathrm{b},1}(\zeta^k_3) = \gamma^k_{\mathrm{b},1}, \;\text{ and }\; \gamma^{\mathrm{b},2}(\zeta^k_3) = \gamma^k_{\mathrm{b},2} \quad\text{on } \partial\omega, \tag{3.13b}$$

(see Table 8 for expressions of the right-hand sides $\boldsymbol{R}^k_{\mathrm{m}}$ and R^k_{b}) where the boundary operators $\gamma^{\mathrm{m},j}$ and $\gamma^{\mathrm{b},j}$, $j=1,2$, depend on the nature of lateral boundary conditions according to Table 2.

3.2.4. Boundary data for $\boldsymbol{\zeta}^0$. For conditions ①–④, the boundary data $\gamma^0_{\mathrm{m},j}$ and $\gamma^0_{\mathrm{b},j}$, $j=1,2$, are all zero, whereas for conditions ⑤–⑧, only the membrane boundary data $\gamma^0_{\mathrm{m},j}$, $j=1,2$, are always zero.

In the cases ⑤ and ⑦, we assume for simplicity that ω is simply connected. Then $\gamma^0_{\mathrm{b},1}$, which is the trace of ζ^0_3 on $\partial\omega$, is a prescribed constant (so that ζ^0_3 has a zero mean value in accordance with the orthogonality condition (2.11)) which is given by the scalar product of R^0_{b} versus the solution of a typical problem of type (3.13b). The other boundary data $\gamma^0_{\mathrm{b},2}$ is zero.

Table 2
Boundary operators for the Kirchhoff–Love generators.

	Membrane part		Bending part	
	$\gamma^{\mathrm{m},1}(\boldsymbol{\zeta}_*)$	$\gamma^{\mathrm{m},2}(\boldsymbol{\zeta}_*)$	$\gamma^{\mathrm{b},1}(\zeta_3)$	$\gamma^{\mathrm{b},2}(\zeta_3)$
①	ζ_n	ζ_s	ζ_3	$\partial_n\zeta_3$
②	ζ_n	$T^{\mathrm{m}}_s(\boldsymbol{\zeta}_*)$	ζ_3	$\partial_n\zeta_3$
③	$T^{\mathrm{m}}_n(\boldsymbol{\zeta}_*)$	ζ_s	ζ_3	$M_n(\zeta_3)$
④	$T^{\mathrm{m}}_n(\boldsymbol{\zeta}_*)$	$T^{\mathrm{m}}_s(\boldsymbol{\zeta}_*)$	ζ_3	$M_n(\zeta_3)$
⑤	ζ_n	ζ_s	ζ_3	$\partial_n\zeta_3$
⑥	ζ_n	$T^{\mathrm{m}}_s(\boldsymbol{\zeta}_*)$	$\partial_n\zeta_3$	$N_n(\zeta_3)$
⑦	$T^{\mathrm{m}}_n(\boldsymbol{\zeta}_*)$	ζ_s	ζ_3	$M_n(\zeta_3)$
⑧	$T^{\mathrm{m}}_n(\boldsymbol{\zeta}_*)$	$T^{\mathrm{m}}_s(\boldsymbol{\zeta}_*)$	$M_n(\zeta_3)$	$N_n(\zeta_3)$

Table 3
Boundary data for $\boldsymbol{\zeta}^1$.

	Membrane part		Bending part	
	$\gamma^1_{\mathrm{m},1}$	$\gamma^1_{\mathrm{m},2}$	$\gamma^1_{\mathrm{b},1}$	$\gamma^1_{\mathrm{b},2}$
①	$c_1^{\textcircled{1}}\,\mathrm{div}_*\,\boldsymbol{\zeta}^0_*$	0	0	$c_4^{\textcircled{1}}\,\Delta_*\zeta^0_3$
②	$c_1^{\textcircled{2}}\,\mathrm{div}_*\,\boldsymbol{\zeta}^0_*$	$c_2^{\textcircled{2}}\,\partial_s\,\mathrm{div}_*\,\boldsymbol{\zeta}^0_*$	0	$c_4^{\textcircled{2}}\,\Delta_*\zeta^0_3$
③	$c_1^{\textcircled{3}}\,\kappa^2\,\zeta^0_n$	0	0	$c_4^{\textcircled{3}}\,\kappa^2\,\partial_n\zeta^0_3$
④	$c_1^{\textcircled{4}}\,\kappa\,\mathrm{div}_*\,\boldsymbol{\zeta}^0_*$	$c_2^{\textcircled{4}}\,\partial_s\,\mathrm{div}_*\,\boldsymbol{\zeta}^0_*$	0	$(c_4^{\textcircled{4}}\,\kappa^2 + c_5^{\textcircled{4}}\,\partial_{ss})\,\partial_n\zeta^0_3$

In the cases ⑥ and ⑧ the boundary condition related to $\gamma^{\mathrm{b},2} = N_n$ is given by

$$N_n(\zeta^0_3) = \frac{3}{2}\left(\int_{-1}^{+1} x_3\, f_n\, dx_3 + g^+_n + g^-_n\right)\bigg|_{\partial\omega}\,. \tag{3.14}$$

The mechanical interpretation of the right-hand side in this relation reads that this expression has the dimension of a moment and can be understood as a prescribed moment on the lateral side of the plate, generated by f_n, g^+_n and g^-_n. Obviously, this right-hand side is zero, if the supports of the data f_n and $g^{\pm}_n$ avoid $\overline{\Gamma}_0$ and $\partial\omega$, respectively. The other boundary data $\gamma^0_{\mathrm{b},1}$ is zero.

3.2.5. Boundary data for $\boldsymbol{\zeta}^1$. For conditions ①–④, all the boundary data for $\boldsymbol{\zeta}^1$ are special traces of $\boldsymbol{\zeta}^0$, according to Table 3 (we recall that κ is the curvature of $\partial\omega$).

Here, the constants $c_j^{\textcircled{i}}$ depend only on λ and μ and come from typical boundary layer profiles.

In contrast to the four "clamped" lateral conditions, for the four "free" lateral conditions ⑤–⑧ the boundary conditions related to the membrane part $\boldsymbol{\zeta}^1_*$ are all zero, which combined with the fact that the interior right-hand side $\boldsymbol{R}^1_{\mathrm{m}}$ is zero, yields that $\boldsymbol{\zeta}^1_*$ is itself zero.

TABLE 4
Boundary data for ζ_3^1.

	$\gamma^1_{\mathrm{b},1}$	$\gamma^1_{\mathrm{b},2}$
⑤	$\Lambda^{\textcircled{5}}$	0
⑥	0	$P^{\textcircled{6}}(\zeta_3^0) + \kappa K^{\textcircled{6}}\Big(f_n, g_n^{\pm}\Big)$
⑦	$\Lambda^{\textcircled{7}}$	$c_4^{\textcircled{7}}\, L$
⑧	$c_3^{\textcircled{8}}\, \partial_s(\partial_n+\kappa)\partial_s\zeta_3^0$	$P^{\textcircled{8}}(\zeta_3^0) + \kappa K^{\textcircled{8}}\Big(f_n, g_n^{\pm}\Big)$

TABLE 5
Typical boundary layer profiles.

Components	$\bar{\boldsymbol{\varphi}}^{\mathrm{m}}$	$\bar{\boldsymbol{\varphi}}^{\mathrm{b}}$	$\bar{\boldsymbol{\varphi}}^{\mathrm{s}}$
Normal	even	odd	0
Horizontal tangential	0	0	odd
Vertical	odd	even	0

The traces of ζ_3^1 are generically not zero: in cases ⑤ and ⑦ (and if ω is simply connected) all traces can be expressed with the help of the function

$$L(s) = \left[-\frac{2}{3}(\tilde{\lambda}+2\mu)\partial_n\Delta_*\,\zeta_3^0 + \int_{-1}^{+1} x_3\, f_n\, dx_3 + g_n^+ + g_n^-\right]\Bigg|_{\partial\omega} . \tag{3.15}$$

In cases ⑥ and ⑧ the prescribed values of the traces involve more complicated operators. We write the boundary data for ζ_3^1 in a condensed form in Table 4.

Here $\Lambda^{\textcircled{5}}$ and $\Lambda^{\textcircled{7}}$ are special double primitives of L on $\partial\omega$. $P^{\textcircled{6}}$ is a linear combination of $\partial_s\kappa^2\partial_s$, $(\kappa\partial_s)^2$ and $\kappa\partial_n\Delta_*$, $P^{\textcircled{8}}$ of $\kappa\partial_n\Delta_*$, $\partial_s(\kappa(\partial_n{+}\kappa))\partial_s$, and $\kappa\partial_s(\partial_n{+}\kappa)\partial_s$. Finally, $K^{\textcircled{6}}$ and $K^{\textcircled{8}}$ are operators preserving the support with respect to the in-plane variables.

3.3. Specific features: The first boundary layer profile. For conditions ①–④, the first boundary layer profile $\boldsymbol{\varphi}^1$ can be described as a sum of three terms in tensor product form in the variables s and (t, x_3) with $t = \frac{r}{\varepsilon}$:

$$\boldsymbol{\varphi}^1 = \ell^{\mathrm{m}}(s)\,\bar{\boldsymbol{\varphi}}^{\mathrm{m}}(t,x_3) + \ell^{\mathrm{b}}(s)\,\bar{\boldsymbol{\varphi}}^{\mathrm{b}}(t,x_3) + \ell^{\mathrm{s}}(s)\,\bar{\boldsymbol{\varphi}}^{\mathrm{s}}(t,x_3) . \tag{3.16}$$

Here $\bar{\boldsymbol{\varphi}}^{\mathrm{m}}$, $\bar{\boldsymbol{\varphi}}^{\mathrm{b}}$, and $\bar{\boldsymbol{\varphi}}^{\mathrm{s}}$ are typical profiles only depending on the Lamé constants and whose components have special parities with respect to x_3: $\bar{\boldsymbol{\varphi}}^{\mathrm{m}}$ is a membrane displacement whereas $\bar{\boldsymbol{\varphi}}^{\mathrm{b}}$ and $\bar{\boldsymbol{\varphi}}^{\mathrm{s}}$ are bending displacements; moreover, some of their components are zero, which is summarized in Table 5.

The functions ℓ are given as traces of $\boldsymbol{\zeta}^0$ along the boundary $\partial\omega$ according to Table 6.

Again in contrast to the four "clamped" lateral conditions, the normal and transverse components of the first boundary layer profile $\boldsymbol{\varphi}^1$ are always zero in the cases ⑤–⑧. Only the in-plane tangential component φ_s^1 is generically nonzero, and it is odd with respect to x_3. This means that $\boldsymbol{\varphi}^1$ is a bending displacement.

Table 6
Lateral traces coming up in the first boundary layer profile.

Case	ℓ^{m}	ℓ^{b}	ℓ^{s}
① and ②	$\mathrm{div}_* \, \zeta^0_*$	$\Delta_* \zeta^0_3$	0
③	$\kappa \, \zeta^0_n$	$\kappa \, \partial_n \zeta^0_3$	0
④	$\mathrm{div}_* \, \zeta^0_*$	$\kappa \, \partial_n \zeta^0_3$	$\partial_s(\partial_n \zeta^0_3)$

Table 7
The first boundary layer profile.

Case	ℓ^{s}	$\bar\varphi^{\mathrm{s}}$
⑤	$\partial_s \zeta^0_3$	$\bar\varphi^{\mathrm{s}}_{\mathrm{Dir}}$
⑥	$\kappa \partial_s \zeta^0_3$	$\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}$
⑦	$\partial_s \zeta^0_3$	$\bar\varphi^{\mathrm{s}}_{\mathrm{Dir}}$
⑧	$(\partial_n + \kappa) \partial_s \zeta^0_3$	$\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}$

The component φ^1_s can be written in tensor product form $\ell^{\mathrm{s}}(s)\, \bar\varphi^{\mathrm{s}}(t, x_3)$ according to Table 7.

Here $\bar\varphi^{\mathrm{s}}_{\mathrm{Dir}}$ and $\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}$ are solutions on the half strip $\mathbb{R}^+ \times (-1, 1)$ of special boundary problems for the Laplace operator; see Lemmas 6.5 and 6.7.

Note the presence of κ in front of the traces for the hard simple support case ③ and for the sliding edge case ⑥ (compare also with [2] and [27], respectively): due to the possibility of reflecting the solution across any flat part of the boundary, the existence of boundary layer terms is linked to nonzero curvature.

4. Inner–outer expansion Ansatz.

4.1. The Ansatz. The determination of the asymptotics (3.2) can be split into two steps. The first one consists of finding *all* suitable power series

$$\underline{\boldsymbol{u}}(\varepsilon)(x) \simeq \underline{\boldsymbol{u}}^0(x) + \varepsilon \underline{\boldsymbol{u}}^1(x) + \cdots + \varepsilon^k \underline{\boldsymbol{u}}^k(x) + \cdots \tag{4.1}$$

which solve in the sense of asymptotic expansions the interior equations (2.12a) in Ω and conditions (2.13a) of traction on the horizontal sides $\Gamma_\pm$. We refer to Maz'ya, Nazarov and Plamenevskii [19, Chap. 15] for general developments relating to the structure of expansion (4.1).

We will see in what follows that all the terms in the suitable series (4.1) are strictly determined *except* the elliptic traces of the Kirchhoff–Love generators $\boldsymbol{\zeta}^k$. The second step, which we will initiate in the next section, consists of finding the profiles $\boldsymbol{w}^k$ so that $\sum_k \varepsilon^k \boldsymbol{w}^k(r\varepsilon^{-1}, s, x_3)$ solves RefPb in Ω with zero volume force, conditions (2.13a) of zero traction, and so that the lateral boundary conditions (2.14) are satisfied by the complete Ansatz. The outcome will be that the existence of *exponentially decaying* profiles is subordinated to the determination of the remaining degrees of freedom in the series (4.1).

4.2. The algorithms of the outer expansion part. This section is devoted to the construction of the most general power series (4.1) solving (2.12a)–(2.13a). Let us introduce the two operators A and B which associate with a displacement $\boldsymbol{u}$ in Ω, a volume force in Ω, and tractions on the horizontal sides on $\Gamma_\pm$ according to

$$A\boldsymbol{u} = \Big(2\mu\,\partial_3 e_{\alpha 3}(\boldsymbol{u}) + \lambda\,\partial_{\alpha 3}u_3,\ (\lambda+2\mu)\partial_{33}u_3\ ;\ 2\mu\, e_{\alpha 3}(\boldsymbol{u})\big|_{\Gamma_\pm},\ (\lambda+2\mu)\partial_3 u_3\big|_{\Gamma_\pm}\Big),$$

$$B\boldsymbol{u} = \Big((\lambda+\mu)\partial_\alpha \mathrm{div}_*\boldsymbol{u}_* + \mu\,\Delta_* u_\alpha,\ \lambda\,\partial_3\,\mathrm{div}_*\boldsymbol{u}_* + 2\mu\,\partial_\beta e_{\beta 3}(\boldsymbol{u})\,;\ 0\big|_{\Gamma_\pm},\ \lambda\,\mathrm{div}_*\boldsymbol{u}_*\big|_{\Gamma_\pm}\Big),$$

the first group of arguments being the in-plane volume forces, the second, the transverse volume force, and similarly for the tractions. Solving (2.12a)–(2.13a) by a power series (4.1) is equivalent to solving the system of equations

$$\tag{4.2} \begin{cases} A\underline{\boldsymbol{u}}^k &= \ 0 \qquad\qquad\qquad \text{for } k=0,1,\\ A\underline{\boldsymbol{u}}^2 + B\underline{\boldsymbol{u}}^0 &= \ \Big(-f_\alpha\ ,\ 0\ ;\ \overset{\pm}{g_\alpha}\big|_{\Gamma_\pm}\ ,\ 0\big|_{\Gamma_\pm}\Big),\\ A\underline{\boldsymbol{u}}^4 + B\underline{\boldsymbol{u}}^2 &= \ \Big(0\ ,\ -f_3\ ;\ 0\big|_{\Gamma_\pm}\ ,\ \overset{\pm}{g_3}\big|_{\Gamma_\pm}\Big),\\ A\underline{\boldsymbol{u}}^k + B\underline{\boldsymbol{u}}^{k-2} &= \ 0 \qquad\qquad\qquad \text{for } k=3 \text{ and } k\ge 5. \end{cases}$$

It is well known that the solutions of the problem $A\underline{\boldsymbol{u}} = 0$ are the Kirchhoff–Love displacements. Thus $\underline{\boldsymbol{u}}^0 = \boldsymbol{u}^0_{\mathrm{KL}}$ and $\underline{\boldsymbol{u}}^1 = \boldsymbol{u}^1_{\mathrm{KL}}$, with generators $\boldsymbol{\zeta}^0$ and $\boldsymbol{\zeta}^1$.

In order to solve the series of equations of *odd order* $A\underline{\boldsymbol{u}}^k + B\underline{\boldsymbol{u}}^{k-2} = 0$, let us introduce the operator V.

DEFINITION 4.1. *The operator* $V : \boldsymbol{\zeta} \mapsto V\boldsymbol{\zeta}$ *is defined from* $\mathcal{C}^\infty(\overline{\omega})^3$ *into* $\mathcal{C}^\infty(\overline{\Omega})^3$ *by*

$$\tag{4.3} \begin{aligned} (V\boldsymbol{\zeta})_\alpha &= \bar p_2\,\partial_\alpha\,\mathrm{div}_*\,\boldsymbol{\zeta}_* + \bar p_3\,\partial_\alpha\Delta_*\zeta_3,\\ (V\boldsymbol{\zeta})_3 &= \bar p_1\ \mathrm{div}_*\,\boldsymbol{\zeta}_* + \bar p_2\,\Delta_*\zeta_3 \end{aligned}$$

with $\bar p_j$ *for* $j=1,2,3$ *the polynomials in the variable* x_3 *of degrees* j *defined as*

$$\tag{4.4} \begin{aligned} \bar p_1(x_3) = -\,\frac{\tilde\lambda}{2\mu}\,x_3, \qquad \bar p_2(x_3) = \frac{\tilde\lambda}{4\mu}\,\Big(x_3^2 - \frac{1}{3}\Big),\\ \bar p_3(x_3) = \frac{1}{12\mu}\,\big((\tilde\lambda+4\mu)\,x_3^3 - (5\tilde\lambda+12\mu)\,x_3\big). \end{aligned}$$

Here $\tilde\lambda$ *still denotes the "homogenized" Lamé coefficient* $2\lambda\mu(\lambda+2\mu)^{-1}$.

With L^{m} the membrane operator (3.8), direct computations yield the following lemma.

LEMMA 4.2. *Let* $\boldsymbol{\zeta}$ *belong to* $\mathcal{C}^\infty(\overline{\omega})^3$ *and let* $\boldsymbol{u}_{\mathrm{KL}}$ *be the associated Kirchhoff–Love displacement. Then the field* $V\boldsymbol{\zeta}$ *is the unique solution with zero mean values on each fiber* $x_*\times(-1,1)$ *of the problem*

$$\tag{4.5} A(V\boldsymbol{\zeta}) + B(\boldsymbol{u}_{\mathrm{KL}}) = \Big(L^{\mathrm{m}}\boldsymbol{\zeta}_*\ ,\ 0\ ;\ 0\big|_{\Gamma_\pm}\ ,\ 0\big|_{\Gamma_\pm}\Big).$$

Then, if $L^{\mathrm{m}}\boldsymbol{\zeta}^1_* = 0$, we can take $\underline{\boldsymbol{u}}^3 = \boldsymbol{u}^3_{\mathrm{KL}} + V\boldsymbol{\zeta}^1$. In order to proceed, we remark that each component of $B(V\boldsymbol{\zeta})$ can be split into two parts, both of them being the product of a polynomial in x_3 and of $\Delta_*\,\mathrm{div}_*\,\boldsymbol{\zeta}_*$ or $\Delta_*^2\zeta_3$, or a derivative of these expressions. With the bending operator (3.11) we easily obtain that if $L^{\mathrm{m}}\boldsymbol{\zeta}_*$ and $L^{\mathrm{b}}\zeta_3$ are zero, then $B(V\boldsymbol{\zeta})$ is zero, too. Thus, the odd part of the outer Ansatz is solved, since we obtain by an induction argument the following proposition.

Proposition 4.3. *For any $k = 1, 3, 5, \ldots$ let $\boldsymbol{\zeta}^k$ be such that $L^{\mathrm{m}}\boldsymbol{\zeta}^k_* = 0$ and $L^{\mathrm{b}}\zeta^k_3 = 0$. Then, setting for $k = 3, 5, \ldots$*

$$\underline{\boldsymbol{u}}^k = \boldsymbol{u}^k_{\mathrm{KL}} + V\boldsymbol{\zeta}^{k-2}\,,$$

we obtain all the solutions of the odd order equations in system (4.2).

Let us consider now the equations of *even* order. The operator A is block triangular and its diagonal is made of ordinary Neumann problems on the interval $(-1, 1)$. So actually, in order to have solvability for these problems, compatibility conditions are required on the right-hand sides. Conversely, if the problems are solvable, the solutions are uniquely determined if we require that they have a mean value zero on each fiber $x_* \times (-1, 1)$ with $x_* \in \overline{\omega}$.

With $\underline{\boldsymbol{u}}^0 = \boldsymbol{u}^0_{\mathrm{KL}}$, we will find $\underline{\boldsymbol{u}}^2$ being of the form $\boldsymbol{u}^2_{\mathrm{KL}} + V\boldsymbol{\zeta}^0 + G(\boldsymbol{f}, \boldsymbol{g}^{\pm})$, where G is another solution operator. But prior to this, we need two sorts of primitives of an integrable function u on the interval $(-1, +1)$.

Notation 4.4. *Let us introduce the following:*

- *the primitive of u with zero mean value on* $(-1, +1)$

$$\oint^{x_3} u\, dy_3 \;:=\; \int_{-1}^{x_3} u(y_3)\, dy_3 - \frac{1}{2}\int_{-1}^{+1}\int_{-1}^{z_3} u(y_3)\, dy_3\, dz_3;$$

- *the primitive of u which vanishes in -1 and 1 if u has a zero mean value on $(-1, +1)$ and which is even, respectively, odd, if u is odd, respectively, even*

$$\fint^{y_3} u\, dz_3 \;:=\; \frac{1}{2}\left(\int_{-1}^{y_3} u(z_3)\, dz_3 - \int_{y_3}^{+1} u(z_3)\, dz_3\right).$$

Definition 4.5. *The operator $G : (\boldsymbol{f}, \boldsymbol{g}^{\pm}) \mapsto G(\boldsymbol{f}, \boldsymbol{g}^{\pm})$ is defined from $\mathcal{C}^\infty(\overline{\Omega})^3 \times \mathcal{C}^\infty(\overline{\omega})^6$ into $\mathcal{C}^\infty(\overline{\Omega})^3$ by*

$$\begin{cases} (G(\boldsymbol{f}, \boldsymbol{g}^{\pm}))_3 &=& 0, \\ (G(\boldsymbol{f}, \boldsymbol{g}^{\pm}))_\alpha &=& \dfrac{1}{2\mu}\displaystyle\oint^{x_3}\left[-2\fint^{y_3} f_\alpha \;+\; \left(g^+_\alpha - g^-_\alpha + \int_{-1}^{+1} f_\alpha\right) y_3 + g^+_\alpha + g^-_\alpha\right] dy_3\,. \end{cases}$$

The reason for the introduction of G is as follows.

Lemma 4.6. *For any $(\boldsymbol{f}, \boldsymbol{g}^{\pm}) \in \mathcal{C}^\infty(\overline{\Omega})^3 \times \mathcal{C}^\infty(\overline{\omega})^6$, $G(\boldsymbol{f}, \boldsymbol{g}^{\pm})$ is the unique solution with zero mean values on each fiber $x_* \times (-1, 1)$ of the problem*

$$A\big(G(\boldsymbol{f}, \boldsymbol{g}^{\pm})\big) = \left(-f_\alpha + \frac{1}{2}\left[\int_{-1}^{+1} f_\alpha\, dx_3 + g^+_\alpha - g^-_\alpha\right]\;,\; 0\;;\; g^{\pm}_\alpha\big|_{\Gamma_\pm}\;,\; 0\big|_{\Gamma_\pm}\right).$$

Now, we can see that if we set

$$\boldsymbol{R}^0_{\mathrm{m}}(x_*) = -\frac{1}{2}\left[\int_{-1}^{+1} \boldsymbol{f}_*(x_*, x_3)\, dx_3 + \boldsymbol{g}^+_*(x_*) - \boldsymbol{g}^-_*(x_*)\right] \tag{4.6}$$

for any $\boldsymbol{\zeta}^0_*$ satisfying the membrane equation $L^{\mathrm{m}}(\boldsymbol{\zeta}^0_*) = \boldsymbol{R}^0_{\mathrm{m}}$, the displacement $\underline{\boldsymbol{u}}^2 = \boldsymbol{u}^2_{\mathrm{KL}} + V\boldsymbol{\zeta}^0 + G(\boldsymbol{f}, \boldsymbol{g}^{\pm})$ solves the equation of order $k = 2$ of system (4.2). We denote by $\boldsymbol{v}^2 = V\boldsymbol{\zeta}^0 + G$ its part with zero mean values on each fiber $x_* \times (-1, 1)$.

In order to go further in solving the even part of the Ansatz, we are going to introduce a residual operator $F = (F_*, F_3)$ and a new solution operator W.

Definition 4.7. (i) *The operator* $F : \boldsymbol{v} \mapsto F\boldsymbol{v} = (F_*\boldsymbol{v}, F_3\boldsymbol{v})$ *is defined from* $\mathcal{C}^\infty(\overline{\Omega})^3$ *into* $\mathcal{C}^\infty(\overline{\omega})^3$ *by*

$$\left\{\begin{array}{rcl} F_3\boldsymbol{v} &=& \mu \displaystyle\int_{-1}^{+1} \partial_\beta e_{\beta 3}(\boldsymbol{v})\, dy_3\,, \\ F_\alpha\boldsymbol{v} &=& \dfrac{\tilde\lambda}{2} \displaystyle\int_{-1}^{+1} \fint^{y_3} \partial_{\alpha\beta} e_{\beta 3}(\boldsymbol{v})\, dz_3\, dy_3\,. \end{array}\right.$$

(ii) *The operator* $W : \boldsymbol{v} \mapsto W\boldsymbol{v}$ *is defined from* $\mathcal{C}^\infty(\overline{\Omega})^3$ *into itself by*

$$\left\{\begin{array}{rcl} W_3\boldsymbol{v} &=& -\displaystyle\oint^{x_3} \Big(\frac{\tilde\lambda}{2\mu}\operatorname{div}_* \boldsymbol{v}_* + \frac{\tilde\lambda}{\lambda}\fint^{y_3} \partial_\beta e_{\beta 3}(\boldsymbol{v})\Big) dy_3, \\ W_\alpha\boldsymbol{v} &=& -\displaystyle\oint^{x_3} \Big(\partial_\alpha W_3\boldsymbol{v} + \fint^{y_3}\Big(\frac{\lambda}{\mu}\,\partial_{\alpha 3} W_3\boldsymbol{v} + \frac{\lambda+\mu}{\mu}\,\partial_\alpha \operatorname{div}_* \boldsymbol{v}_* + \Delta_* v_\alpha\Big)\Big) dy_3. \end{array}\right.$$

With these operators, we can prove the following lemma.

Lemma 4.8. *Let* $\boldsymbol{v}$ *in* $\mathcal{C}^\infty(\overline{\Omega})^3$ *be a displacement field with zero mean values on each fiber* $x_* \times (-1, 1)$, $x_* \in \overline{\omega}$. *Then* $W\boldsymbol{v}$ *has also zero mean values on each fiber* $x_* \times (-1, 1)$ *and solves the problem*

$$A(W\boldsymbol{v}) + B(\boldsymbol{v}) = \Big(0\ ,\ 0\ ;\ \pm F_*(\boldsymbol{v})\big|_{\Gamma_\pm}\ ,\ \mp F_3(\boldsymbol{v})\big|_{\Gamma_\pm}\Big).$$

Now, it is natural to search for $\underline{\boldsymbol{u}}^4$ with the form $\boldsymbol{u}^4_{\mathrm{KL}} + V\boldsymbol{\zeta}^2 + W(V\boldsymbol{\zeta}^0 + G) + H$. In view of Lemmas 4.2 and 4.8, with such an Ansatz, H has to solve the problem

$$AH = \Big(-L^{\mathrm{m}}(\boldsymbol{\zeta}^2_*)\ ,\ -f_3\ ;\ \mp F_*(V\boldsymbol{\zeta}^0 + G)\big|_{\Gamma_\pm}\ ,\ g_3^\pm \pm F_3(V\boldsymbol{\zeta}^0 + G)\big|_{\Gamma_\pm}\Big). \tag{4.7}$$

Thus, it is important to have more information about $F(V\boldsymbol{\zeta}^0 + G)$. It is not difficult to check, as seen below.

Lemma 4.9. *For all* $\boldsymbol{\zeta}$ *in* $\mathcal{C}^\infty(\overline{\omega})^3$ *we have*

$$F_*(V\boldsymbol{\zeta}) = 0 \qquad \textit{and} \qquad F_3(V\boldsymbol{\zeta}) = -\tfrac{1}{3}L^{\mathrm{b}}\zeta_3\,.$$

Moreover, we have

$$F_3(G) = \frac{1}{2}\operatorname{div}_* \Big[\int_{-1}^{+1} x_3\, \boldsymbol{f}_*\, dx_3 + \boldsymbol{g}_*^+ + \boldsymbol{g}_*^-\Big]\,. \tag{4.8}$$

Then there holds the following lemma.

Lemma 4.10. *Let* R^0_{b} *be defined as*

$$R^0_{\mathrm{b}} = \frac{3}{2}\Big[\int_{-1}^{+1} f_3\, dx_3 + g_3^+ - g_3^- + \operatorname{div}_*\Big(\int_{-1}^{+1} x_3\, \boldsymbol{f}_*\, dx_3 + \boldsymbol{g}_*^+ + \boldsymbol{g}_*^-\Big)\Big] \tag{4.9}$$

and $\boldsymbol{R}^2_{\mathrm{m}}$ *be defined as*

$$\boldsymbol{R}^2_{\mathrm{m}} = F_*(G) - \frac{\tilde\lambda}{4\mu}\nabla_*\Big[\int_{-1}^{+1} x_3\, f_3\, dx_3 + g_3^+ + g_3^-\Big]\,. \tag{4.10}$$

TABLE 8
Algorithm formulas.

k	$\underline{\boldsymbol{u}}^k$	$\boldsymbol{v}^k$	$\boldsymbol{y}^{k-2}$	$\boldsymbol{R}^k_{\mathrm{m}}$	R^k_{b}
0	$\boldsymbol{u}^0_{\mathrm{KL}}$	—	—	$\boldsymbol{R}^0_{\mathrm{m}}$	R^0_{b}
2	$\boldsymbol{u}^2_{\mathrm{KL}}+\boldsymbol{v}^2$	$V\boldsymbol{\zeta}^0+\boldsymbol{y}^0$	G	$\boldsymbol{R}^2_{\mathrm{m}}$	$3F_3(W\boldsymbol{v}^2+H)$
4	$\boldsymbol{u}^4_{\mathrm{KL}}+\boldsymbol{v}^4$	$V\boldsymbol{\zeta}^2+\boldsymbol{y}^2$	$W\boldsymbol{v}^2+H$	$F_*\boldsymbol{v}^4$	$3F_3(W\boldsymbol{v}^4+Y\boldsymbol{\zeta}^4_*)$
$2\ell{+}2$	$\boldsymbol{u}^{2\ell+2}_{\mathrm{KL}}+\boldsymbol{v}^{2\ell+2}$	$V\boldsymbol{\zeta}^{2\ell}+\boldsymbol{y}^{2\ell}$	$W\boldsymbol{v}^{2\ell}+Y\boldsymbol{\zeta}^{2\ell}_*$	$F_*\boldsymbol{v}^{2\ell+2}$	$3F_3(W\boldsymbol{v}^{2\ell+2}{+}Y\boldsymbol{\zeta}^{2\ell+2}_*)$
1	$\boldsymbol{u}^1_{\mathrm{KL}}$	—	—	0	0
$2\ell{+}1$	$\boldsymbol{u}^{2\ell+1}_{\mathrm{KL}}+\boldsymbol{v}^{2\ell+1}$	$V\boldsymbol{\zeta}^{2\ell-1}$	—	0	0

If there hold $L^{\mathrm{b}}(\zeta^0_3)=R^0_{\mathrm{b}}$ *and* $L^{\mathrm{m}}(\boldsymbol{\zeta}^2_*)=\boldsymbol{R}^2_{\mathrm{m}}$, *then* (4.7) *admits a unique solution* $H=H(\boldsymbol{f},\boldsymbol{g}^{\pm})$ *with zero mean values on each fiber* $x_*\times(-1,1)$ *which is given by*

$$\left\{\begin{array}{lcl} H_3 &=& \dfrac{1}{2(\lambda+2\mu)}\displaystyle\oint^{x_3}\Big[\Big(-2\,{-}\!\!\!\!\!\!\int^{y_3} f_3\Big)+g^+_3+g^-_3\Big]dy_3, \\ H_\alpha &=& -\displaystyle\oint^{x_3}\Big[\partial_\alpha H_3+\frac{1}{\mu}\,y_3\,F_*(G)+\frac{\lambda}{\mu}\,{-}\!\!\!\!\!\!\int^{y_3}\Big\{\partial_{\alpha3}H_3-\frac{1}{2}\int_{-1}^{+1}\partial_{\alpha3}H_3\,dz_3\Big\}\Big]dy_3\,. \end{array}\right.$$

Thus, we have found $\underline{\boldsymbol{u}}^4$ as $\boldsymbol{u}^4_{\mathrm{KL}}+\boldsymbol{v}^4$, where $\boldsymbol{v}^4$ has zero mean values on each fiber $x_*\times(-1,1)$: $\boldsymbol{v}^4$ is given by $V\boldsymbol{\zeta}^2+W(V\boldsymbol{\zeta}^0+G)+H=V\boldsymbol{\zeta}^2+W\boldsymbol{v}^2+H$.

Next, we search for a $\underline{\boldsymbol{u}}^6$ with the form $\boldsymbol{u}^6_{\mathrm{KL}}+V\boldsymbol{\zeta}^4+W\boldsymbol{v}^4+Y$. In view of Lemmas 4.2 and 4.8, with such an Ansatz, Y has to solve the problem

$$AY=\Big(-L^{\mathrm{m}}(\boldsymbol{\zeta}^4_*)\;,\;0\;;\;\mp F_*(\boldsymbol{v}^4)\big|_{\Gamma_\pm}\;,\;\pm F_3(\boldsymbol{v}^4)\big|_{\Gamma_\pm}\Big). \tag{4.11}$$

This problem is solvable if (i) $F_3(\boldsymbol{v}^4)$ is zero, which holds true if $L^{\mathrm{b}}\zeta^2_3=3F_3(W\boldsymbol{v}^2{+}H)$, (ii) $L^{\mathrm{m}}(\boldsymbol{\zeta}^4_*)=F_*(\boldsymbol{v}^4)$, compare Lemma 4.10.

Then $Y=Y(\boldsymbol{\zeta}^4_*)$ solves (4.11), with the solution operator Y defined as follows.

DEFINITION 4.11. *For* $\boldsymbol{\zeta}_*\in\mathcal{C}^\infty(\overline{\omega})^2$, $Y=Y(\boldsymbol{\zeta}_*)$ *is defined as*

$$Y_3=0\quad\textit{and}\quad Y_*=-2\,\tilde{\lambda}^{-1}\,\bar{p}_2\,L^{\mathrm{m}}(\boldsymbol{\zeta}_*)\,.$$

From now on, the solution of (4.2) is self-similar. Summarizing, we obtain by induction that every expansion (4.1) solving (2.12a)–(2.13a) can be described according to Table 8, where G and H are a condensed notation for $G(\boldsymbol{f},\boldsymbol{g}^{\pm})$ and $H(\boldsymbol{f},\boldsymbol{g}^{\pm})$, respectively, and $\boldsymbol{R}^k_{\mathrm{m}}$ and R^k_{b} are the prescribed values for $L^{\mathrm{m}}(\boldsymbol{\zeta}^k_*)$ and $L^{\mathrm{b}}(\zeta^k_3)$, respectively (note that $\boldsymbol{R}^0_{\mathrm{m}}$, $\boldsymbol{R}^2_{\mathrm{m}}$, and R^0_{b} are defined in (4.6), (4.10), and (4.9)).

Here the even order terms and the odd order ones are independent from each other. We will see later on that they are connected by the lateral boundary conditions via the boundary layer terms. We emphasize that each term $\underline{\boldsymbol{u}}^k$ in the algorithm is the sum of two terms $\underline{\boldsymbol{u}}^k=\boldsymbol{u}^k_{\mathrm{KL}}+\boldsymbol{v}^k$ with $\boldsymbol{u}^k_{\mathrm{KL}}$ representing the general solution of homogeneous Neumann problems for ordinary differential equations over each fiber $x_*\times(-1,1)$ and $\boldsymbol{v}^k$ being particular solutions of inhomogeneous ordinary Neumann problems across the thickness with mean value zero.

4.3. Formulas for the determined part of the displacements. The formulas in Table 8 giving the $\boldsymbol{v}^k$ yield in a straightforward way that

$$\begin{aligned} \boldsymbol{v}^{2k+1} &= V\boldsymbol{\zeta}^{2k-1}\,, \\ \boldsymbol{v}^{2k+2} &= \sum_{\ell=0}^{k} W^\ell\circ V\boldsymbol{\zeta}^{2(k-\ell)} + \sum_{\ell=0}^{k-2} W^\ell\circ Y\boldsymbol{\zeta}_*^{2(k-\ell)} \\ &\quad + W^k\circ G(\boldsymbol{f},\boldsymbol{g}^\pm) + W^{k-1}\circ H(\boldsymbol{f},\boldsymbol{g}^\pm) \end{aligned} \tag{4.12}$$

with the convention that $W^{-1}=0$ and $W^0=\mathrm{Id}$.

Using the definitions of V and W, we can prove the following.

LEMMA 4.12. *For $\ell=0,1,\ldots,$ we have the following formulas for the iterates $W^\ell\circ V$:*

$$\begin{aligned} (W^\ell\circ V\boldsymbol{\zeta})_\alpha &= \bar r_{2\ell+2}\,\partial_\alpha\Delta_*^\ell\operatorname{div}_*\boldsymbol{\zeta}_* + \bar r_{2\ell+3}\,\partial_\alpha\Delta_*^{\ell+1}\zeta_3, \\ (W^\ell\circ V\boldsymbol{\zeta})_3 &= \bar q_{2\ell+1}\,\Delta_*^\ell\operatorname{div}_*\boldsymbol{\zeta}_* + \bar q_{2\ell+2}\,\Delta_*^{\ell+1}\zeta_3 \end{aligned} \tag{4.13}$$

with $\bar q_j$, $\bar r_j$ the polynomials in the variable x_3 of degrees j and of parities j defined recursively as

$$\bar q_1=\bar p_1,\quad \bar q_2=\bar p_2,\quad \bar r_2=\bar p_2,\quad \bar r_3=\bar p_3,$$

with $\bar p_j$ for $j=1,2,3$, the polynomials defined in (4.4)*, and*

$$\begin{aligned} \bar q_j(x_3) &= -\oint^{x_3}\Big(\frac{\tilde\lambda}{2\mu}\bar r_{j-1} + \frac{\tilde\lambda}{2\lambda}\fint^{y_3}(\bar q_{j-2}+\bar r'_{j-1})\Big)dy_3 && \textit{for } j\ge 3, \\ \bar r_j(x_3) &= -\oint^{x_3}\Big(\bar q_{j-1} + \fint^{y_3}\Big(\frac{\lambda}{\mu}\,\bar q'_{j-1} + \frac{\lambda+2\mu}{\mu}\,\bar r_{j-2}\Big)\Big)dy_3, && \textit{for } j\ge 4. \end{aligned} \tag{4.14}$$

Similarly, using the definition of Y, we are able to show the following lemma.

LEMMA 4.13. *For $\ell=0,1,\ldots,$ we have the following formulas for the iterates $W^\ell\circ Y$:*

$$\begin{aligned} (W^\ell\circ Y\boldsymbol{\zeta}_*)_\alpha &= \bar s_{2\ell+2}\,\partial_\alpha\Delta_*^\ell\operatorname{div}_*\boldsymbol{\zeta}_* + \bar t_{2\ell+2}\,\Delta_*^{\ell+1}\zeta_\alpha, \\ (W^\ell\circ Y\boldsymbol{\zeta}_*)_3 &= \bar s_{2\ell+1}\,\Delta_*^\ell\operatorname{div}_*\boldsymbol{\zeta}_* \end{aligned} \tag{4.15}$$

with $\bar s_j$ and $\bar t_j$ the polynomials in the variable x_3 of degrees j and of parities j defined recursively as

$$\bar s_1=0,\quad \bar s_2=-\,\frac{\lambda+2\mu}{\lambda}\,\bar p_2\quad \textit{and}\quad \bar t_2=-\,\frac{3\lambda+2\mu}{\lambda}\,\bar p_2$$

with $\bar p_2$ given in (4.4)*, and for $\ell\ge 1$*

$$\begin{aligned} \bar s_{2\ell+1}(x_3) &= -\oint^{x_3}\Big(\frac{\tilde\lambda}{2\mu}(\bar s_{2\ell}+\bar t_{2\ell}) + \frac{\tilde\lambda}{2\lambda}\fint^{y_3}(\bar s_{2\ell-1}+\bar s'_{2\ell}+\bar t'_{2\ell})\Big)dy_3, \\ \bar s_{2\ell+2}(x_3) &= -\oint^{x_3}\Big(\bar s_{2\ell+1} + \fint^{y_3}\Big(\frac{\lambda}{\mu}\,\bar s'_{2\ell+1} + \frac{\lambda+\mu}{\mu}\,(\bar s_{2\ell}+\bar t_{2\ell}) + \bar s_{2\ell}\Big)\Big)dy_3, \\ \bar t_{2\ell+2}(x_3) &= -\oint^{x_3}\Big(\fint^{y_3}\bar t_{2\ell}\Big)dy_3. \end{aligned} \tag{4.16}$$

Condensing $G(\boldsymbol{f}, \boldsymbol{g}^{\pm})$ into G and $H(\boldsymbol{f}, \boldsymbol{g}^{\pm})$ into H, we obtain the following formulas for the first $\boldsymbol{v}^k$ (k even):

$$\begin{aligned} v_\alpha^2 &= \bar{p}_2\, \partial_\alpha \operatorname{div}_* \boldsymbol{\zeta}_*^0 + \bar{p}_3\, \partial_\alpha \Delta_* \zeta_3^0 + G_\alpha, \\ v_3^2 &= \bar{p}_1 \operatorname{div}_* \boldsymbol{\zeta}_*^0 + \bar{p}_2\, \Delta_* \zeta_3^0\,, \end{aligned} \tag{4.17}$$

$$\begin{aligned} v_\alpha^4 &= \bar{p}_2\, \partial_\alpha \operatorname{div}_* \boldsymbol{\zeta}_*^2 + \bar{p}_3\, \partial_\alpha \Delta_* \zeta_3^2 + \bar{r}_4\, \partial_\alpha \Delta_* \operatorname{div}_* \boldsymbol{\zeta}_*^0 + \bar{r}_5\, \partial_\alpha \Delta_*^2 \zeta_3^0 + (WG + H)_\alpha, \\ v_3^4 &= \bar{p}_1 \operatorname{div}_* \boldsymbol{\zeta}_*^2 + \bar{p}_2\, \Delta_* \zeta_3^2 + \bar{q}_3\, \Delta_* \operatorname{div}_* \boldsymbol{\zeta}_*^0 + \bar{q}_4\, \Delta_*^2 \zeta_3^0 + (WG + H)_3\,. \end{aligned} \tag{4.18}$$

5. The principles of construction of the inner expansion part.

After the construction of the most general power series (4.1) solving (2.12a)–(2.13a), we see that the only remaining degrees of freedom can be given by traces of the Kirchhoff–Love generators $\boldsymbol{\zeta}^k$. As will be investigated for each case in particular, complementing traces of the Kirchhoff–Love generators $\boldsymbol{\zeta}^k$ can be determined along with the computation of the boundary layer terms $\boldsymbol{w}^k$.

The boundary layer Ansatz, namely $\sum_{k\geq 1} \varepsilon^k \boldsymbol{w}^k$, must satisfy the equations (2.12a) inside Ω with vanishing right-hand side and the boundary conditions (2.13a) of zero traction on the horizontal faces of Ω, and must compensate for the lateral boundary conditions of the power series $\sum_{k\geq 0} \varepsilon^k \underline{\boldsymbol{u}}^k$ so that the lateral boundary conditions (2.14) are fulfilled. We present in this section some common features of all problems.

5.1. The equations of the inner expansion.

5.1.1. Lateral boundary conditions. In order to obtain the relations which have to be satisfied by the inner part of the expansion, we evaluate the boundary conditions for a displacement $\boldsymbol{u}$ of the form

$$\boldsymbol{u}(\varepsilon)(x) = \underline{\boldsymbol{u}}(x) + (\boldsymbol{\varphi}_*, \varepsilon\varphi_3)\Big(\frac{r}{\varepsilon}, s, x_3\Big), \tag{5.1}$$

where $\underline{\boldsymbol{u}} = \sum_{k\geq 0} \varepsilon^k \underline{\boldsymbol{u}}^k$ and $\boldsymbol{\varphi} = \sum_{k\geq 1} \varepsilon^k \boldsymbol{\varphi}^k$. The form of the boundary layer term $(\boldsymbol{\varphi}_*, \varepsilon\varphi_3)$ is related to the covariant nature of displacements: indeed we return with $\boldsymbol{\varphi}$ to the homogeneity of the original unknown $\boldsymbol{u}^\varepsilon$. We denote by φ_t the normal component of $\boldsymbol{\varphi}$.

For $\boldsymbol{u}$ of the form (5.1), the formulas for the lateral Dirichlet conditions are obvious, and the lateral Neumann conditions can be written with the help of the following boundary operators acting on the profiles $\boldsymbol{\varphi}$:

$$\begin{aligned} T_t^{(0)}(\boldsymbol{\varphi}) &= \lambda\, \partial_3 \varphi_3 + (\lambda + 2\mu)\partial_t \varphi_t, & T_t^{(1)}(\boldsymbol{\varphi}) &= \lambda(\partial_s \varphi_s - \tfrac{1}{R}\, \varphi_t), \\ T_s^{(0)}(\boldsymbol{\varphi}) &= \mu\, \partial_t \varphi_s, & T_s^{(1)}(\boldsymbol{\varphi}) &= \mu(\partial_s \varphi_t + \tfrac{2}{R}\, \varphi_s), \\ T_3^{(0)}(\boldsymbol{\varphi}) &= \mu(\partial_t \varphi_3 + \partial_3 \varphi_t). \end{aligned} \tag{5.2}$$

Thus, we can write the components of the lateral traction (cf. (2.15a)) as

$$T_n(\varepsilon) = \varepsilon\, T_t^{(0)}(\boldsymbol{\varphi}) + \varepsilon^2 T_t^{(1)}(\boldsymbol{\varphi}) + \lambda\, \partial_3\, \underline{u}_3 + \varepsilon^2\big(\lambda \operatorname{div}_* \underline{\boldsymbol{u}}_* + 2\mu\, \partial_n \underline{u}_n\big), \tag{5.3a}$$

$$T_s(\varepsilon) = \varepsilon\, T_s^{(0)}(\boldsymbol{\varphi}) + \varepsilon^2 T_s^{(1)}(\boldsymbol{\varphi}) + \varepsilon^2 \mu\big(\partial_s \underline{u}_n + \partial_n \underline{u}_s + \tfrac{2}{R}\, \underline{u}_s\big), \tag{5.3b}$$

$$T_3(\varepsilon) = T_3^{(0)}(\boldsymbol{\varphi}) + \mu(\partial_n \underline{u}_3 + \partial_3 \underline{u}_n). \tag{5.3c}$$

5.1.2. Interior equations. In variables (t, s, x_3) and unknowns

$$\boldsymbol{\varphi} = (\varphi_t, \varphi_s, \varphi_3) \sim \Big(\boldsymbol{w}_*, \frac{1}{\varepsilon}\, w_3\Big),$$

the interior equations (2.12a) for $\boldsymbol{w}$ have the form

$$\mathcal{B}(\varepsilon\,;\, t, s\,;\, \partial_t, \partial_s, \partial_3)\boldsymbol{\varphi} = 0,$$

where the three components $\mathcal{B}(\varepsilon)_t$, $\mathcal{B}(\varepsilon)_s$, and $\mathcal{B}(\varepsilon)_3$ of $\mathcal{B}(\varepsilon)$ can be written as polynomials of degree 2 in ε with coefficients involving partial derivative operators of degree ≤ 2 combined with integer powers of $R = R(s)$ and of $\frac{1}{\rho}$ with

$$\rho = R(s) - r = R(s) - \varepsilon t$$

which is the curvature radius in s of the curve $\{x_* \in \omega,\ \operatorname{dist}(x_*, \partial\omega) = r\}$. The thorough expression of $\mathcal{B}(\varepsilon)$ can be found in [11, section 3]. A Taylor expansion at $t = 0$ of $\rho^{-1} = (R - \varepsilon t)^{-1}$ yields an asymptotic expansion of $\mathcal{B}$ in a power series of ε:

$$\mathcal{B} \sim \mathcal{B}^{(0)} + \varepsilon\mathcal{B}^{(1)} + \cdots \varepsilon^k \mathcal{B}^{(k)} + \cdots, \tag{5.4}$$

where the $\mathcal{B}^{(k)}(t, s\,;\, \partial_t, \partial_s, \partial_3)$ are partial differential systems of order 2 with polynomial coefficients in t independent from ε. Here follow the expressions for $\mathcal{B}^{(0)}$ and $\mathcal{B}^{(1)}$:

$$\begin{aligned}
(\mathcal{B}^{(0)}\boldsymbol{\varphi})_t &= \mu\big(\partial_{tt}\varphi_t + \partial_{33}\varphi_t\big) + (\lambda+\mu)\,\partial_t\big(\partial_t\varphi_t + \partial_3\varphi_3\big),\\
(\mathcal{B}^{(0)}\boldsymbol{\varphi})_s &= \mu\big(\partial_{tt}\varphi_s + \partial_{33}\varphi_s\big),\\
(\mathcal{B}^{(0)}\boldsymbol{\varphi})_3 &= \mu\,\big(\partial_{tt}\varphi_3 + \partial_{33}\varphi_3\big) + (\lambda+\mu)\,\partial_3\big(\partial_t\varphi_t + \partial_3\varphi_3\big)
\end{aligned} \tag{5.5}$$

and, with the curvature $\kappa = \frac{1}{R}$,

$$\begin{aligned}
(\mathcal{B}^{(1)}\boldsymbol{\varphi})_t &= -\mu\,\kappa\,\partial_t\varphi_t + (\lambda+\mu)\,\partial_t\big(-\kappa\,\varphi_t + \partial_s\varphi_s\big),\\
(\mathcal{B}^{(1)}\boldsymbol{\varphi})_s &= \mu\,\kappa\big(\partial_{tt}(t\varphi_s) + \partial_{33}(t\varphi_s)\big) - \mu\,\kappa\,\partial_t\varphi_s + (\lambda+\mu)\,\partial_s\big(\partial_t\varphi_t + \partial_3\varphi_3\big),\\
(\mathcal{B}^{(1)}\boldsymbol{\varphi})_3 &= -\mu\,\kappa\,\partial_t\varphi_3 + (\lambda+\mu)\,\partial_3\big(-\kappa\,\varphi_t + \partial_s\varphi_s\big).
\end{aligned} \tag{5.6}$$

Thus, the interior equation $\mathcal{B}(\varepsilon)\boldsymbol{\varphi} = 0$ can be written as

$$\mathcal{B}^{(0)}\boldsymbol{\varphi} + \varepsilon\mathcal{B}^{(1)}\boldsymbol{\varphi} + \cdots \varepsilon^k\mathcal{B}^{(k)}\boldsymbol{\varphi} + \cdots \sim 0. \tag{5.7}$$

5.1.3. Horizontal boundary conditions. The boundary conditions on the horizontal sides $x_3 = \pm 1$ are (cf. (2.13a))

$$\mu(\partial_3\varphi_t + \partial_t\varphi_3) = 0, \tag{5.8a}$$

$$\mu\partial_3\varphi_s + \varepsilon\,\mu\,\partial_s\varphi_3 = 0, \tag{5.8b}$$

$$(\lambda + 2\mu)\partial_3\varphi_3 + \lambda\,\partial_t\varphi_t + \varepsilon\,\lambda\bigg(-\tfrac{1}{\rho}\varphi_t + \tfrac{R}{\rho}\partial_s\Big(\tfrac{R}{\rho}\varphi_s\Big)\bigg) = 0. \tag{5.8c}$$

Similarly to the interior equations, we can develop the horizontal boundary conditions $\mathcal{G}$ (5.8a) in powers of ε:

$$\mathcal{G} \sim \mathcal{G}^{(0)} + \varepsilon\mathcal{G}^{(1)} + \cdots \varepsilon^k\mathcal{G}^{(k)} + \cdots, \tag{5.9}$$

where the $\mathcal{G}^{(k)}(t, s\,;\partial_t, \partial_s, \partial_3)$ are partial differential systems of order 1 with polynomial coefficients in t. The expressions for $\mathcal{G}^{(0)}$ and $\mathcal{G}^{(1)}$ are

$$(5.10)\quad \begin{array}{lcllcl} (\mathcal{G}^{(0)}\boldsymbol{\varphi})_t &=& \mu(\partial_3\varphi_t + \partial_t\varphi_3), & (\mathcal{G}^{(1)}\boldsymbol{\varphi})_t &=& 0, \\ (\mathcal{G}^{(0)}\boldsymbol{\varphi})_s &=& \mu\partial_3\varphi_s, & (\mathcal{G}^{(1)}\boldsymbol{\varphi})_s &=& \mu\partial_s\varphi_3, \\ (\mathcal{G}^{(0)}\boldsymbol{\varphi})_3 &=& (\lambda+2\mu)\partial_3\varphi_3 + \lambda\,\partial_t\varphi_t, & (\mathcal{G}^{(1)}\boldsymbol{\varphi})_3 &=& \lambda(-\kappa\,\varphi_t + \partial_s\varphi_s). \end{array}$$

Thus, the horizontal boundary conditions $\mathcal{G}(\varepsilon)\boldsymbol{\varphi} = 0$ can be written as

$$(5.11)\qquad \mathcal{G}^{(0)}\boldsymbol{\varphi} + \varepsilon\mathcal{G}^{(1)}\boldsymbol{\varphi} + \cdots \varepsilon^k\mathcal{G}^{(k)}\boldsymbol{\varphi} + \cdots \sim 0.$$

5.2. The recursive equations. Assuming that $\sum_k \varepsilon^k \underline{\boldsymbol{u}}^k$ already fulfills the relations in Table 8, we determine now the equations satisfied by the profiles $\boldsymbol{\varphi}^k$ and the remaining conditions satisfied by the displacements $\underline{\boldsymbol{u}}^k$ so that

$$(5.12)\qquad \sum_{k\,\geq\,0} \varepsilon^k \underline{\boldsymbol{u}}^k + \sum_{k\,\geq\,1} \varepsilon^k (\boldsymbol{\varphi}_*^k, \varepsilon\varphi_3^k)$$

satisfies equations (2.12a)–(2.14).

5.2.1. Interior equations. Equation (5.7) yields that

$$(5.13)\qquad \forall k \geq 0, \qquad \sum_{\ell=0}^{k} \mathcal{B}^{(\ell)}\boldsymbol{\varphi}^{k-\ell} = 0,$$

which guarantees (2.12a) for the whole expansion (5.12).

5.2.2. Horizontal boundary conditions. Equation (5.11) yields that

$$(5.14)\qquad \forall k \geq 0, \qquad \sum_{\ell=0}^{k} \mathcal{G}^{(\ell)}\boldsymbol{\varphi}^{k-\ell} = 0,$$

which guarantees (2.13a) for the whole expansion (5.12).

5.2.3. Lateral Dirichlet boundary conditions. Let $\sum_k \varepsilon^k D_n^k$, $\sum_k \varepsilon^k D_s^k$, and $\sum_k \varepsilon^k D_3^k$ be the normal, tangential, and vertical components of the lateral Dirichlet traces of the series (5.12). The lateral Dirichlet boundary conditions then read

$$(5.15)\quad \forall k \geq 0, \qquad D_n^k = 0 \text{ if } n \in A, \quad D_s^k = 0 \text{ if } s \in A, \quad D_3^k = 0 \text{ if } 3 \in A,$$

which immediately yields the Dirichlet conditions for the whole expansion (5.12).

For the terms D^k, we have

$$(5.16)\qquad D_n^0 = \underline{u}_n^0, \quad D_s^0 = \underline{u}_s^0, \quad D_3^0 = \underline{u}_3^0, \quad D_3^1 = \underline{u}_3^1,$$

and for $k \geq 1$

$$(5.17a)\qquad D_n^k = \varphi_t^k + \underline{u}_n^k,$$

$$(5.17b)\qquad D_s^k = \varphi_s^k + \underline{u}_s^k,$$

$$(5.17c)\qquad D_3^{k+1} = \varphi_3^k + \underline{u}_3^{k+1}\,.$$

5.2.4. Lateral Neumann boundary conditions. Let $\sum_k \varepsilon^k T_n^k$, $\sum_k \varepsilon^k T_s^k$, and $\sum_k \varepsilon^k T_3^k$ be the normal, tangential, and vertical components of the lateral Neumann traces of the series (5.12). The lateral Neumann boundary conditions then read

$$\forall k \geq 0, \qquad T_n^k = 0 \text{ if } n \in B, \quad T_s^k = 0 \text{ if } s \in B, \quad T_3^k = 0 \text{ if } 3 \in B, \tag{5.18}$$

which immediately yields the Neumann conditions for the whole expansion (5.12).

Let us evaluate the terms T^k. To that aim, we rely on the following formulas for $\underline{\boldsymbol{u}}^k$, cf. Table 8. Either $\underline{\boldsymbol{u}}^k = \boldsymbol{u}^k_{\mathrm{KL}} + \boldsymbol{v}^k$, i.e.,

$$\underline{u}_n^k = \zeta_n^k - x_3\,\partial_n \zeta_3^k + v_n^k, \tag{5.19a}$$

$$\underline{u}_s^k = \zeta_s^k - x_3\,\partial_s \zeta_3^k + v_s^k, \tag{5.19b}$$

$$\underline{u}_3^k = \zeta_3^k + v_3^k, \tag{5.19c}$$

or $\underline{\boldsymbol{u}}^k = \boldsymbol{u}^k_{\mathrm{KL}} + V\boldsymbol{\zeta}^{k-2} + \boldsymbol{y}^{k-2}$, i.e.,

$$\underline{u}_n^k = \zeta_n^k - x_3\,\partial_n \zeta_3^k + \bar{p}_2\,\partial_n \operatorname{div}_* \boldsymbol{\zeta}_*^{k-2} + \bar{p}_3\,\partial_n \Delta_* \zeta_3^{k-2} + \boldsymbol{y}_n^{k-2}, \tag{5.20a}$$

$$\underline{u}_s^k = \zeta_s^k - x_3\,\partial_s \zeta_3^k + \bar{p}_2\,\partial_s \operatorname{div}_* \boldsymbol{\zeta}_*^{k-2} + \bar{p}_3\,\partial_s \Delta_* \zeta_3^{k-2} + \boldsymbol{y}_s^{k-2}, \tag{5.20b}$$

$$\underline{u}_3^k = \zeta_3^k + \bar{p}_1\, \operatorname{div}_* \boldsymbol{\zeta}_*^{k-2} + \bar{p}_2\,\Delta_* \zeta_3^{k-2} + \boldsymbol{y}_3^{k-2}, \tag{5.20c}$$

where $\bar{p}_1$, $\bar{p}_2$, $\bar{p}_3$ are introduced in (4.4).

Thus, we find

$$T_n^0 = 0, \quad T_n^1 = 0, \quad T_s^0 = 0, \quad T_s^1 = 0, \quad T_3^0 = 0, \tag{5.21}$$

and for $k \geq 1$ (cf. (3.10a), (3.12a), (5.2))

$$\begin{aligned} {}^{k+1}_{n} = {} & T_t^{(0)}(\boldsymbol{\varphi}^k) + T_t^{(1)}(\boldsymbol{\varphi}^{k-1}) + T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^{k-1}) - x_3\, M_n(\zeta_3^{k-1}) \\ & + \lambda\,\partial_3\, \boldsymbol{y}_3^{k-1} + \lambda \operatorname{div}_* \boldsymbol{v}_*^{k-1} + 2\mu\,\partial_n v_n^{k-1}, \end{aligned} \tag{5.22a}$$

$$\begin{aligned} T_s^{k+1} = {} & T_s^{(0)}(\boldsymbol{\varphi}^k) + T_s^{(1)}(\boldsymbol{\varphi}^{k-1}) + T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^{k-1}) - 2\mu x_3\,(\partial_n + \tfrac{1}{R})\partial_s \zeta_3^{k-1} \\ & + \mu\big(\partial_s v_n^{k-1} + \partial_n v_s^{k-1} + \tfrac{2}{R}\, v_s^{k-1}\big), \end{aligned} \tag{5.22b}$$

$$\begin{aligned} T_3^k = {} & T_3^{(0)}(\boldsymbol{\varphi}^k) + \mu(\bar{p}_2 + \bar{p}_3')\,\partial_n \Delta_* \zeta_3^{k-2} \\ & + \mu(\partial_n \boldsymbol{y}_3^{k-2} + \partial_3 \boldsymbol{y}_n^{k-2}). \end{aligned} \tag{5.22c}$$

5.3. Solving the inner expansion. According to the calculations of the previous subsection, to solve the problem with the Ansatz (5.12), it remains to find a sequence of profiles $(\boldsymbol{\varphi}^k)_k$ and a sequence of Kirchhoff–Love generators $(\boldsymbol{\zeta}^k)_k$ such that (5.13), (5.14), (5.15), and (5.18) hold.

Let us consider now the profiles $\boldsymbol{\varphi}^k$ for $k \geq 1$ as main unknowns. In view of (5.13), (5.14), (5.17a), and (5.22), we see that the sequence of problems satisfied by the $\boldsymbol{\varphi}^k$ can be written in a recursive way: for each $k \geq 1$ the profile $\boldsymbol{\varphi}^k$ has to solve the equation

$$\mathcal{B}_{\textcircled{\scriptsize i}}(\boldsymbol{\varphi}^k) = (\mathfrak{f}^k; \mathfrak{g}^k; \mathfrak{h}^k), \tag{5.23}$$

where

- $\mathcal{B}_{\textcircled{i}}$ is the operator $\mathcal{B}^{(0)}$ inside the domain, the traction operator $\mathcal{G}^{(0)}$ on the horizontal sides, the Dirichlet traces on the lateral side for $a \in A_{\textcircled{i}}$ and the Neumann traces on the lateral side for $b \in B_{\textcircled{i}}$,
- $\mathfrak{f}^k$ and $\mathfrak{g}^k$ are the following functions of the previous profiles

$$\mathfrak{f}^k = -\sum_{\ell=1}^{k} \mathcal{B}^{(\ell)} \boldsymbol{\varphi}^{k-\ell} \quad \text{and} \quad \mathfrak{g}^k = -\sum_{\ell=1}^{k} \mathcal{G}^{(\ell)} \boldsymbol{\varphi}^{k-\ell}, \tag{5.24}$$

so that (5.13)–(5.14) is solved, and $\mathfrak{h}^k$ involves previous profiles as well as certain traces of the Kirchhoff–Love generators, according to (5.15)–(5.22).

An important point is now to note that neither $\mathcal{B}^{(0)}$, nor $\mathcal{G}^{(0)}$, nor the lateral trace operators of $\mathcal{B}_{\textcircled{i}}$ contain any derivative with respect to the tangential variable s. Thus, (5.23) can be solved in the variables $t \in \mathbb{R}^+$ and $x_3 \in (-1,1)$, the role of s being only that of a parameter. So we introduce the half-strip

$$\Sigma^+ = \big\{(t, x_3); \quad 0 < t, \quad -1 < x_3 < 1\big\}. \tag{5.25}$$

Its boundary has two horizontal parts $\gamma_\pm = \mathbb{R}^+ \times \{x_3 = \pm 1\}$ and a lateral part

$$\gamma_0 = \big\{(t, x_3); \quad t = 0, \quad -1 < x_3 < 1\big\}. \tag{5.26}$$

Thus, we have

$$\mathcal{B}_{\textcircled{i}}(\boldsymbol{\varphi}) = (\mathfrak{f}; \mathfrak{g}; \mathfrak{h}) \quad \Longleftrightarrow \quad \left\{ \begin{array}{rcll} \mathcal{B}^{(0)}(\boldsymbol{\varphi}) & = & \mathfrak{f}, & \text{in } \Sigma^+, \\ \mathcal{G}^{(0)}(\boldsymbol{\varphi}) & = & \mathfrak{g}, & \text{on } \gamma_\pm, \\ \varphi_a & = & \mathfrak{h}_a, & \text{on } \gamma_0, \quad \forall a \in A_{\textcircled{i}}, \\ T_b^{(0)}(\boldsymbol{\varphi}) & = & \mathfrak{h}_b, & \text{on } \gamma_0, \quad \forall b \in B_{\textcircled{i}}. \end{array} \right. \tag{5.27}$$

Essential is the possibility of finding *exponentially decreasing* solutions when $\mathfrak{f}$ and $\mathfrak{g}$ have the same property. This is what we start to investigate in the next section.

6. Exponentially decaying profiles in a half-strip.

6.1. General principles. The properties of the operators $\mathcal{B}_{\textcircled{i}}$ are closely linked to those of the corresponding operator $\underline{\mathcal{B}}$ on the full strip $\Sigma := \mathbb{R} \times (-1,1)$, defined as $\underline{\mathcal{B}}(\boldsymbol{\varphi}) = (\mathfrak{f}; \mathfrak{g})$ with $\mathfrak{f} = \mathcal{B}^{(0)}(\boldsymbol{\varphi})$ in Σ and $\mathfrak{g} = \mathcal{G}^{(0)}(\boldsymbol{\varphi})$ on $\mathbb{R} \times \{x_3 = \pm 1\}$; see also Nazarov and Plamenevskii [23, Chap. 5].

Let $\mathcal{P}$ be the space of polynomial displacements $\boldsymbol{Z}$ satisfying $\underline{\mathcal{B}}(\boldsymbol{Z}) = 0$. Computations like those of Mielke in [20] yield that $\mathcal{P}$ has eight dimensions and that a basis of $\mathcal{P}$ is given by the following polynomial displacements $\boldsymbol{Z}^{[1]}, \dots, \boldsymbol{Z}^{[8]}$

$$\boldsymbol{Z}^{[1]} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{Z}^{[2]} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \boldsymbol{Z}^{[3]} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \boldsymbol{Z}^{[4]} = \begin{pmatrix} -x_3 \\ 0 \\ t \end{pmatrix},$$

$$\boldsymbol{Z}^{[5]} = \begin{pmatrix} t \\ 0 \\ \bar{p}_1 \end{pmatrix}, \qquad \boldsymbol{Z}^{[6]} = \begin{pmatrix} 0 \\ t \\ 0 \end{pmatrix}, \qquad \boldsymbol{Z}^{[7]} = \begin{pmatrix} -2tx_3 \\ 0 \\ t^2 + 2\bar{p}_2 \end{pmatrix},$$

$$\boldsymbol{Z}^{[8]} = \begin{pmatrix} -3t^2 x_3 + 6\bar{p}_3 \\ 0 \\ t^3 + 6t\bar{p}_2 \end{pmatrix},$$

where $\bar{p}_1(x_3)$, $\bar{p}_2(x_3)$, $\bar{p}_3(x_3)$ are the polynomials previously introduced in (4.4).

Let us introduce weighted spaces H^m_η on the half-strip Σ^+: for $\eta > 0$, their elements are exponentially decreasing as $t \to \infty$.

Definition 6.1. *Let $\eta \in \mathbb{R}$. For $m \geq 0$ let $H^m_\eta(\Sigma^+)$ be the space of functions v such that $e^{\eta t} v$ belongs to $H^m(\Sigma^+)$. We also denote $H^0_\eta(\Sigma^+)$ by $L^2_\eta(\Sigma^+)$. Similar definitions hold for $\mathbb{R}^+$.*

Like in [9, Lemmas 4.10 and 4.11], we have, with η_0 the smallest exponent arising from the Papkovich–Fadle eigenfunctions, compare Papkovich [26] for early reference and Gregory and Wan [17]:

Lemma 6.2. *Let η, $0 < \eta < \eta_0$. Let $\mathfrak{f}$ belong to $L^2_\eta(\Sigma^+)^3$ and $\mathfrak{g}$ belong to $L^2_\eta(\mathbb{R}^+)^6$, let $\mathfrak{h}_a$ belong to $H^{1/2}(\gamma_0)$ for each $a \in A_{\textcircled{i}}$, and $\mathfrak{h}_b$ belong to $H^{-1/2}(\gamma_0)$ for each $b \in B_{\textcircled{i}}$. Then there exist $\boldsymbol{\varphi} \in H^1_\eta(\Sigma^+)^3$ and $\boldsymbol{Z} \in \mathcal{P}$ so that*

$$\mathcal{B}_{\textcircled{i}}(\boldsymbol{\varphi} + \boldsymbol{Z}) = (\mathfrak{f}; \mathfrak{g}; \mathfrak{h}). \tag{6.1}$$

But the solution given by Lemma 6.2 is not unique. Let $\mathcal{T}_{\textcircled{i}}$ denote the space of the polynomial displacements $\boldsymbol{Z}$ such that there exists $\boldsymbol{\varphi} = \boldsymbol{\varphi}(\boldsymbol{Z}) \in H^1_\eta(\Sigma^+)^3$ satisfying

$$\mathcal{B}_{\textcircled{i}}(\boldsymbol{Z} + \boldsymbol{\varphi}(\boldsymbol{Z})) = 0.$$

Like in [9, Proposition 4.12], we can prove that the dimension of $\mathcal{T}_{\textcircled{i}}$ is 4. Thus $\mathcal{P}$ can be split in the direct sum of two four-dimensional spaces $\mathcal{Z}_{\textcircled{i}}$ and $\mathcal{T}_{\textcircled{i}}$, and we have as a corollary the following lemma.

Lemma 6.3. *Let $\mathfrak{f}$, $\mathfrak{g}$, and $\mathfrak{h}$ be as in Lemma* 6.2. *Then there exist $\boldsymbol{\varphi}$* unique *in $H^1_\eta(\Sigma^+)^3$ and $\boldsymbol{Z}$* unique *in the four-dimensional space $\mathcal{Z}_{\textcircled{i}}$ so that* (6.1) *holds.*

At this stage, the conclusion is that we have a defect number equal to four for the solution of the sequence of the above equations (5.23) by exponentially decreasing displacements $\boldsymbol{\varphi}^k$ for each $s \in \partial\omega$. But four traces on $\partial\omega$ are still available, allowing us to modify $\mathfrak{h}^k$. Note that this is coherent with the principle of "matching asymptotics," according to which the behavior at infinity of the profiles is transformed into a function of the primitive variable x (which is a Kirchhoff–Love displacement).

6.2. The operators acting on profiles. We can immediately see that the operators $\mathcal{B}_{\textcircled{i}}$ act separately on the couple of components (φ_t, φ_3) that we denote $\boldsymbol{\varphi}_\natural$ and on φ_s. The elasticity operator with the Lamé constants λ and μ acts on $\boldsymbol{\varphi}_\natural$, and the Laplace operatoron φ_s.

The interior elasticity operator in Σ^+ is

$$\mathcal{B}^{(0)}_\natural : \ \boldsymbol{\varphi}_\natural \quad \longmapsto \quad \mathfrak{f}_\natural \ = \ \mu(\partial_{tt} + \partial_{33}) \begin{pmatrix} \varphi_t \\ \varphi_3 \end{pmatrix} + (\lambda + \mu) \begin{pmatrix} \partial_t \\ \partial_3 \end{pmatrix} (\partial_t \varphi_t + \partial_3 \varphi_3), \tag{6.2}$$

its horizontal boundary conditions $\mathcal{G}^{(0)}$ (5.10) on $\gamma_\pm$ are

$$\mathcal{G}^{(0)}_\natural : \ \boldsymbol{\varphi}_\natural \quad \longmapsto \quad \mathfrak{g}_\natural \ = \ \begin{pmatrix} \mu(\partial_3 \varphi_t + \partial_t \varphi_3) \\ (\lambda + 2\mu)\partial_3 \varphi_3 + \lambda\, \partial_t \varphi_t \end{pmatrix}, \tag{6.3}$$

and the lateral boundary conditions are either Dirichlet's or Neumann's acting on the traction $\boldsymbol{T}^{(0)}_\natural = (T^{(0)}_t, T^{(0)}_3)$; cf. (5.25).

Let us introduce the four elasticity operators that we need. For each of them $\mathfrak{f}_\natural = \mathcal{B}^{(0)}_\natural(\boldsymbol{\varphi}_\natural)$ and $\mathfrak{g}_\natural = \mathcal{G}^{(0)}_\natural(\boldsymbol{\varphi}_\natural)$. Only differs the definition of the lateral trace $\mathfrak{h}_\natural$:

- $E_{\rm Dir}$: $\boldsymbol{\varphi}_\natural \mapsto (\mathfrak{f}_\natural;\mathfrak{g}_\natural;\mathfrak{h}_\natural)$ with $\mathfrak{h}_\natural$ the trace of $\boldsymbol{\varphi}_\natural$ on γ_0,
- $E_{\rm Mix1}$: $\boldsymbol{\varphi}_\natural \mapsto (\mathfrak{f}_\natural;\mathfrak{g}_\natural;\mathfrak{h}_\natural)$ with $\mathfrak{h}_\natural$ the trace of $\big(T^{(0)}_t(\boldsymbol{\varphi}_\natural),\varphi_3\big)$ on γ_0,
- $E_{\rm Mix2}$: $\boldsymbol{\varphi}_\natural \mapsto (\mathfrak{f}_\natural;\mathfrak{g}_\natural;\mathfrak{h}_\natural)$ with $\mathfrak{h}_\natural$ the trace of $\big(\varphi_t, T^{(0)}_3(\boldsymbol{\varphi}_\natural)\big)$ on γ_0,
- $E_{\rm Free}$: $\boldsymbol{\varphi}_\natural \mapsto (\mathfrak{f}_\natural;\mathfrak{g}_\natural;\mathfrak{h}_\natural)$ with $\mathfrak{h}_\natural$ the trace of $\boldsymbol{T}^{(0)}_\natural(\boldsymbol{\varphi}_\natural)$ on γ_0,

whereas the Laplace operators are defined as

- $L_{\rm Dir}$: $\varphi_s \mapsto (\mathfrak{f}_s;\mathfrak{g}_s;\mathfrak{h}_s)$ with $\mathfrak{f}_s = \mu\Delta\varphi_s$, $\mathfrak{g}_s = \mu\partial_3\varphi_s$ and $\mathfrak{h}_s = \varphi_s$ on γ_0,
- $L_{\rm Neu}$: $\varphi_s \mapsto (\mathfrak{f}_s;\mathfrak{g}_s;\mathfrak{h}_s)$ with $\mathfrak{f}_s = \mu\Delta\varphi_s$, $\mathfrak{g}_s = \mu\partial_3\varphi_s$ and $\mathfrak{h}_s = \mu\partial_t\varphi_s$ on γ_0.

Then we have the splittings

$$\mathcal{B}_{\textcircled{1}} = E_{\rm Dir}\oplus L_{\rm Dir}, \quad \mathcal{B}_{\textcircled{2}} = E_{\rm Dir}\oplus L_{\rm Neu}, \quad \mathcal{B}_{\textcircled{3}} = E_{\rm Mix1}\oplus L_{\rm Dir}, \quad \mathcal{B}_{\textcircled{4}} = E_{\rm Mix1}\oplus L_{\rm Neu},$$
$$\mathcal{B}_{\textcircled{5}} = E_{\rm Mix2}\oplus L_{\rm Dir}, \quad \mathcal{B}_{\textcircled{6}} = E_{\rm Mix2}\oplus L_{\rm Neu}, \quad \mathcal{B}_{\textcircled{7}} = E_{\rm Free}\oplus L_{\rm Dir}, \quad \mathcal{B}_{\textcircled{8}} = E_{\rm Free}\oplus L_{\rm Neu}.$$

6.3. The Laplacian on the half-strip. The Neumann problem on the full strip Σ has a polynomial kernel of dimension 2 generated by 1 and t, corresponding to the elements $\boldsymbol{Z}^{[2]}$ and $\boldsymbol{Z}^{[6]}$ of the space $\mathcal{P}$ introduced at the beginning of the section.

6.3.1. Operator protect $\boldsymbol{L_{\rm Dir}}$. The polynomial kernel of this problem is the function t and by integration by parts of $t\,\Delta(\varphi+\delta)$ on rectangles $\Sigma_L = (0,L)\times(-1,1)$ with $L\to+\infty$, we easily prove the following proposition.

Proposition 6.4. *For $\eta>0$, let $f\in L^2_\eta(\Sigma^+)$, $g^\pm \in L^2_\eta(\mathbb{R}^+)^2$, and $h\in H^{1/2}(\gamma_0)$. If, moreover, $\eta<\pi/2$, then the problem*

$$L_{\rm Dir}(\psi) = \Big(f; g^\pm; h\Big)$$

has a unique solution $\psi = \varphi+\delta$ in $H^1_\eta(\Sigma^+)\oplus\operatorname{span}\{1\}$ with $\varphi\in H^1_\eta(\Sigma^+)$ and

$$\text{(6.4)}\ \ \delta = \frac{1}{2\mu}\left(-\int_{\Sigma^+} t\,f(t,x_3)\,dt\,dx_3 + \int_{\mathbb{R}^+} t\big(g^+(t)-g^-(t)\big)\,dt + \mu\int_{-1}^{+1} h(x_3)\,dx_3\right).$$

Later on we will use as a model profile the exponentially decaying solution $\bar\varphi^{\rm s}_{\rm Dir}$ of a special problem involving $L_{\rm Dir}$.

Lemma 6.5. *Let $\bar\varphi^{\rm s}_{\rm Dir}\in H^1_\eta(\Sigma^+)$ be the exponentially decaying solution of the problem*

$$L_{\rm Dir}(\bar\varphi^{\rm s}_{\rm Dir}) = (0;0;x_3)\,;$$

then it holds

$$\int_0^\infty \bar\varphi^{\rm s}_{\rm Dir}(t,1)\,dt > 0.$$

Proof. The function $\bar\varphi^{\rm s}_{\rm Dir}$ is an odd function with respect to x_3. Hence $\bar\varphi^{\rm s}_{\rm Dir}(t,0) = 0$ for $t\in\mathbb{R}^+$. Moreover, as $\bar\varphi^{\rm s}_{\rm Dir}$ is harmonic, it can be reflected by parity at the line $x_3=1$ according to the reflection principle of Schwarz for harmonic functions. Thus, we obtain a function $\tilde\varphi$, which is still harmonic, but now in $\widetilde{\Sigma}^+ = \mathbb{R}^+\times(0,2)$. Hence $\tilde\varphi$ satisfies the Dirichlet problem $\Delta\,\tilde\varphi = 0$ in $\widetilde{\Sigma}^+$ and $\tilde\varphi = \tilde\Phi$ on $\partial\widetilde{\Sigma}^+$ with $\Phi(t,x_3)=0$ for $x_3 = 0$, 2, and any t and $\Phi(0,x_3) = x_3$ for $0<x_3\le 1$ and $\Phi(0,x_3) = 2-x_3$ for $1\le x_3<2$. From the maximum principle for harmonic functions it follows $\tilde\varphi>0$ in $\widetilde{\Sigma}^+$, hence the assertion. $\square$

6.3.2. Operator $\boldsymbol{L}_{\mathbf{Neu}}$. The polynomial kernel of this problem is the function 1 and there holds the following proposition similarly.

Proposition 6.6. *For $\eta > 0$, let $f \in L^2_\eta(\Sigma^+)$, $g^\pm \in L^2_\eta(\mathbb{R}^+)^2$, and $h \in H^{-1/2}(\gamma_0)$. If, moreover, $\eta < \pi/2$, then the problem*

$$L_{\mathrm{Neu}}(\psi) = \Big(f; g^\pm; h\Big)$$

has a unique solution $\psi = \varphi + \delta\, t$ in $H^1_\eta(\Sigma^+) \oplus \mathrm{span}\{t\}$ with $\varphi \in H^1_\eta(\Sigma^+)$ and

$$(6.5)\quad \delta = \frac{1}{2\mu}\left(\int_{\Sigma^+} f(t,x_3)\,dt\,dx_3 - \int_{\mathbb{R}^+}\big(g^+(t) - g^-(t)\big)\,dt + \int_{-1}^{+1} h(x_3)\,dx_3\right).$$

We introduce the solution $\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}$ similarly as above, and using the second Green formula for the product $x_3\,\Delta\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}(t,x_3)$ on Σ_+ we prove the following lemma.

Lemma 6.7. *Let $\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}} \in H^1_\eta(\Sigma^+)$ be the exponentially decaying solution of the problem*

$$L_{\mathrm{Neu}}(\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}) = (0; 0; 2\mu x_3)\,;$$

then it holds

$$\int_0^\infty \bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}(t,1)\,dt = -\frac{2}{3}\,.$$

6.4. Elasticity on the half-strip. The problem (6.2)–(6.3) on the full strip has a polynomial kernel of dimension six generated by $\boldsymbol{Z}^{[1]}_\natural, \boldsymbol{Z}^{[3]}_\natural, \boldsymbol{Z}^{[4]}_\natural, \boldsymbol{Z}^{[5]}_\natural, \boldsymbol{Z}^{[7]}_\natural, \boldsymbol{Z}^{[8]}_\natural$, where the two components of $\boldsymbol{Z}^{[j]}_\natural$ are the first and third ones of $\boldsymbol{Z}^{[j]}$. In particular a basis of the two-dimensional rigid motions is given by

$$\boldsymbol{Z}^{[1]}_\natural = \begin{pmatrix}1\\0\end{pmatrix}, \qquad \boldsymbol{Z}^{[3]}_\natural = \begin{pmatrix}0\\1\end{pmatrix}, \qquad \boldsymbol{Z}^{[4]}_\natural = \begin{pmatrix}-x_3\\t\end{pmatrix}.$$

6.4.1. Operator $\boldsymbol{E}_{\mathbf{Dir}}$. From [9, Proposition 4.12], we obtain the following proposition.

Proposition 6.8. *For $\eta > 0$, let $\mathfrak{f}_\natural \in L^2_\eta(\Sigma^+)^2$, $\mathfrak{g}^\pm_\natural \in L^2_\eta(\mathbb{R}^+)^4$, and $\mathfrak{h}_\natural \in H^{1/2}(\gamma_0)^2$. If, moreover, $\eta < \eta_0$, then the problem*

$$E_{\mathrm{Dir}}(\boldsymbol{\psi}) = \Big(\mathfrak{f}_\natural; \mathfrak{g}^\pm_\natural; \mathfrak{h}_\natural\Big)$$

has a unique solution in $H^1_\eta(\Sigma^+)^2 \oplus \mathrm{span}\,\{\boldsymbol{Z}^{[1]}_\natural, \boldsymbol{Z}^{[3]}_\natural, \boldsymbol{Z}^{[4]}_\natural\}$.

6.4.2. Other operators. Concerning the other operators E_{Mix1}, E_{Mix2}, and E_{Free}, and in contrast to the case of E_{Dir}, they have a polynomial kernel generated by some of the $\boldsymbol{Z}^{[j]}_\natural$. Relying on the following duality relations (6.7) satisfied by the $\boldsymbol{Z}^{[j]}$, formulas for the coefficients in the asymptotics at infinity of the solutions can be obtained from integrations by parts.

Lemma 6.9. *Let $\boldsymbol{T}^{(0)}$ denote the lateral inward traction operator $(T_t^{(0)}, T_s^{(0)}, T_3^{(0)})$; cf.* (5.2). *With σ the permutation*

$$\begin{aligned}
&\sigma(1)=5, \quad \sigma(2)=6, \quad \sigma(3)=8, \quad \sigma(4)=7,\\
&\sigma(5)=1, \quad \sigma(6)=2, \quad \sigma(7)=4, \quad \sigma(8)=3,
\end{aligned}$$

the antisymmetrized flux, which can be defined for any $L \in \mathbb{R}$ by

$$\underline{\Phi}\Big(\boldsymbol{Z}^{[i]}, \boldsymbol{Z}^{[j]}\Big) := \int_{-1}^{+1} \Big(\boldsymbol{T}^{(0)}\Big(\boldsymbol{Z}^{[i]}\Big)\cdot \boldsymbol{Z}^{[j]} - \boldsymbol{T}^{(0)}\Big(\boldsymbol{Z}^{[j]}\Big)\cdot \boldsymbol{Z}^{[i]}\Big)(L, x_3)\, dx_3 \tag{6.6}$$

is independent of L (compare [9, *Lemma* 3.1]) *and satisfies, for $i, j \in \{1, \ldots, 8\}$*

$$\underline{\Phi}\Big(\boldsymbol{Z}^{[i]}, \boldsymbol{Z}^{[j]}\Big) = \bar\gamma_i\, \delta_{j\sigma(i)}, \tag{6.7}$$

with $\bar\gamma_i$ a nonzero real number.

For $i = 2, 6$ we find again the simple relations on which rely Propositions 6.4 and 6.6. For the remaining values of i, the relations (6.7) apply to the bidimensional displacements $\boldsymbol{Z}_\natural^{[i]}$. Relying on (6.7) and integration by parts, we are able to present formulas for the coefficients in the asymptotics at infinity of the solutions to the problems concerning the operators E_{Mix1}, E_{Mix2}, and E_{Free}.

Proposition 6.10. *For $\eta > 0$, let $\mathfrak{f}_\natural \in L^2_\eta(\Sigma^+)^2$, $\mathfrak{g}_\natural^{\pm} \in L^2_\eta(\mathbb{R}^+)^4$, $\mathfrak{h}_t \in H^{-1/2}(\gamma_0)$, and $\mathfrak{h}_3 \in H^{1/2}(\gamma_0)$. If, moreover, $\eta < \eta_0$, then the problem*

$$E_{\mathrm{Mix1}}(\boldsymbol{\psi}) = \Big(\mathfrak{f}_\natural; \mathfrak{g}_\natural^{\pm}; \mathfrak{h}_\natural\Big)$$

has a unique solution $\boldsymbol{\psi} = \boldsymbol{\varphi} + \delta_3 \boldsymbol{Z}_\natural^{[3]} + \delta_5 \boldsymbol{Z}_\natural^{[5]} + \delta_7 \boldsymbol{Z}_\natural^{[7]}$ with $\boldsymbol{\varphi} \in H^1_\eta(\Sigma^+)^2$ and

$$\bar\gamma_5 \delta_5 = \int_{\Sigma^+} f_t - \int_{\mathbb{R}^+} (g_t^+ - g_t^-) + \int_{-1}^{+1} \mathfrak{h}_t\,, \tag{6.8a}$$

$$\bar\gamma_7 \delta_7 = \int_{\Sigma^+} (-x_3 f_t + t f_3) + \int_{\mathbb{R}^+} \Big(g_t^+ + g_t^- - t(g_3^+ - g_3^-)\Big) - \int_{-1}^{+1} x_3 \mathfrak{h}_t\,, \tag{6.8b}$$

(6.8c)

$$\bar\gamma_3 \delta_3 = \int_{\Sigma^+} \mathfrak{f}_\natural \cdot \boldsymbol{Z}_\natural^{[8]} - \int_{\mathbb{R}^+} \Big(\mathfrak{g}^+ \cdot \boldsymbol{Z}_\natural^{[8]}\big|_{\gamma^+} - \mathfrak{g}^- \cdot \boldsymbol{Z}_\natural^{[8]}\big|_{\gamma^-}\Big) + 6\int_{-1}^{+1} \bar p_3\, \mathfrak{h}_t - \mu(\bar p_2 + \bar p_3')\mathfrak{h}_3\,.$$

Proposition 6.11. *For $\eta > 0$, let $\mathfrak{f}_\natural \in L^2_\eta(\Sigma^+)^2$, $\mathfrak{g}_\natural^{\pm} \in L^2_\eta(\mathbb{R}^+)^4$, $\mathfrak{h}_t \in H^{1/2}(\gamma_0)$, and $\mathfrak{h}_3 \in H^{-1/2}(\gamma_0)$. If, moreover, $\eta < \eta_0$, then the problem*

$$E_{\mathrm{Mix2}}(\boldsymbol{\psi}) = \Big(\mathfrak{f}_\natural; \mathfrak{g}_\natural^{\pm}; \mathfrak{h}_\natural\Big)$$

has a unique solution $\boldsymbol{\psi} = \boldsymbol{\varphi} + \delta_1 \boldsymbol{Z}_\natural^{[1]} + \delta_4 \boldsymbol{Z}_\natural^{[4]} + \delta_8 \boldsymbol{Z}_\natural^{[8]}$ with $\boldsymbol{\varphi} \in H^1_\eta(\Sigma^+)^2$ and

$$\bar\gamma_8 \delta_8 = \int_{\Sigma^+} f_3 - \int_{\mathbb{R}^+} (g_3^+ - g_3^-) + \int_{-1}^{+1} \mathfrak{h}_3\,, \tag{6.9a}$$

$$\bar\gamma_1\delta_1 = \int_{\Sigma^+} tf_t - \int_{\mathbb{R}^+} t(g_t^+ - g_t^-) - \int_{-1}^{+1} (\tilde\lambda+2\mu)\mathfrak{h}_t - \frac{\tilde\lambda}{2\mu}\Big(\int_{\Sigma^+} x_3 f_3 - \int_{\mathbb{R}^+}(g_3^+ + g_3^-) + \int_{-1}^{+1} x_3\mathfrak{h}_3\Big)\,,$$

(6.9b)

$$\text{(6.9c)}\quad \bar\gamma_4\delta_4 = \int_{\Sigma^+} \mathfrak{f}_\natural\cdot \boldsymbol{Z}_\natural^{[7]} - \int_{\mathbb{R}^+}\Big(\mathfrak{g}^+\cdot\boldsymbol{Z}_\natural^{[7]} - \mathfrak{g}^-\cdot\boldsymbol{Z}_\natural^{[7]}\Big) + 2\int_{-1}^{+1}\Big(\bar p_2\mathfrak{h}_3 + (\tilde\lambda+2\mu)x_3\mathfrak{h}_t\Big)\,.$$

Proposition 6.12. *For $\eta > 0$, let $\mathfrak{f}_\natural \in L^2_\eta(\Sigma^+)^2$, $\mathfrak{g}^{\pm}_\natural \in L^2_\eta(\mathbb{R}^+)^4$, and $\mathfrak{h}_\natural \in H^{-1/2}(\gamma_0)^2$. If, moreover, $\eta < \eta_0$, then the problem*

$$E_{\text{Free}}(\boldsymbol{\psi}) = \Big(\mathfrak{f}_\natural; \overset{\pm}{\mathfrak{g}}_\natural; \mathfrak{h}_\natural\Big)$$

has a unique solution $\boldsymbol{\psi} = \boldsymbol{\varphi} + \delta_5\boldsymbol{Z}_\natural^{[5]} + \delta_7\boldsymbol{Z}_\natural^{[7]} + \delta_8\boldsymbol{Z}_\natural^{[8]}$ with $\boldsymbol{\varphi}\in H^1_\eta(\Sigma^+)^2$ and

$$\text{(6.10a)}\qquad \bar\gamma_5\delta_5 = \int_{\Sigma^+} f_t - \int_{\mathbb{R}^+}(g_t^+ - g_t^-) + \int_{-1}^{+1}\mathfrak{h}_t\,,$$

$$\text{(6.10b)}\qquad \bar\gamma_8\delta_8 = \int_{\Sigma^+} f_3 - \int_{\mathbb{R}^+}(g_3^+ - g_3^-) + \int_{-1}^{+1}\mathfrak{h}_3\,,$$

$$\text{(6.10c)}\quad \bar\gamma_7\delta_7 = \int_{\Sigma^+}(-x_3 f_t + tf_3) + \int_{\mathbb{R}^+}\big(g_t^+ + g_t^- - t(g_3^+ - g_3^-)\big) - \int_{-1}^{+1} x_3\mathfrak{h}_t\,.$$

7. Clamped plates.

7.1. Hard clamped plates: The first terms in the asymptotics. In [19, Chap. 16], Maz'ya, Nazarov, and Plamenevskii prove estimates like (3.4) for isotropic clamped plates and in [8, 9] the analog of Theorem 3.2 is proved for monoclinic clamped plates.

Here we will show how the formulas relating to lateral boundary condition ① in Tables 2, 3, 5, and 6 can be derived.

From (5.16) it follows that the boundary operators for the generators are the Dirichlet ones and that the four traces of $\boldsymbol{\zeta}^0$ are zero. We find again a fact known for long; cf. [5, 13] for early reference.

Let us investigate $\boldsymbol{\zeta}^1$ and $\boldsymbol{\varphi}^1$ simultaneously. Condition (5.15) for $k=1$ yields that $\zeta_3^1 = 0$, $\varphi_n^1 + \zeta_n^1 - x_3\partial_n\zeta_3^1 = 0$ and $\varphi_s^1 + \zeta_s^1 - x_3\partial_s\zeta_3^1 = 0$ on Γ_0. Moreover, condition (5.15) for $k=2$ with (5.17c) yields that $\varphi_3^1 + \zeta_3^2 + v_3^2 = 0$ on Γ_0.

Thus, the first profile $\boldsymbol{\varphi}^1(s) : (t,x_3)\mapsto\boldsymbol{\varphi}^1(t,s,x_3)$ has to solve for all $s\in\partial\omega$ (cf. (5.23)), the equation $\mathcal{B}_{①}(\boldsymbol{\varphi}^1(s)) = (0;0;\mathfrak{h}^1(s))$ with the trace $\mathfrak{h}^1(s)$ equal to

$$\mathfrak{h}_n^1(s) = -(\zeta_n^1 - x_3\partial_n\zeta_3^1)(s),\quad \mathfrak{h}_s^1(s) = -(\zeta_s^1 - x_3\partial_s\zeta_3^1)(s),\quad \mathfrak{h}_3^1(s) = -(\zeta_3^2 + v_3^2)(s).$$

Note that the unknowns are the profile $\boldsymbol{\varphi}^1$ and the traces of ζ_n^1, ζ_s^1, $\partial_n\zeta_3^1$, and ζ_3^2.

Since $\mathcal{B}_{①}$ splits into the direct sum $E_{\text{Dir}}\oplus L_{\text{Dir}}$, for each $s\in\partial\omega$ (fixed now, thus omitted),

- φ_s^1 is the solution of the Poisson problem

$$\text{(7.1)}\qquad L_{\text{Dir}}(\varphi_s^1) = (0;0;\mathfrak{h}_s^1),$$

- the couple $\boldsymbol{\varphi}^1_\natural$ is the solution of the elasticity system

$$E_{\mathrm{Dir}}(\boldsymbol{\varphi}^1_\natural) = (0; 0; \mathfrak{h}^1_\natural). \tag{7.2}$$

We have to find the conditions on $\boldsymbol{\zeta}^1$ so that (7.1) and (7.2) admit exponentially decreasing solutions.

Concerning the Poisson problem, Proposition 6.4 yields that (7.1) admits an exponentially decreasing solution if the coefficient (6.4) is zero, i.e., if $\int_{-1}^{+1} \mathfrak{h}^1_s = 0$. With the above expression of $\mathfrak{h}^1_s$, this yields that $\zeta^1_s = 0$ on $\partial\omega$. Since we already found that $\zeta^1_3 = 0$ on $\partial\omega$, we obtain that $\mathfrak{h}^1_s \equiv 0$, thus $\varphi^1_s = 0$.

Concerning the Lamé problem, Proposition 6.8 yields a solution for (7.2) in $H^1_\eta(\Sigma^+)^2 \oplus \mathrm{span}\{\boldsymbol{Z}^{[1]}_\natural, \boldsymbol{Z}^{[3]}_\natural, \boldsymbol{Z}^{[4]}_\natural\}$. We first recall that (cf. (4.3)–(4.4))

$$v^2_3(x_*, x_3) = \bar{p}_1(x_3)\ \mathrm{div}_*\, \boldsymbol{\zeta}^0_*(x_*) + \bar{p}_2(x_3)\ \Delta_* \zeta^0_3(x_*)\,. \tag{7.3}$$

Let $\bar{\boldsymbol{\psi}}^{\mathrm{m}}_\natural$ be the solution in $H^1_\eta(\Sigma^+)^2 \oplus \mathrm{span}\{\boldsymbol{Z}^{[1]}_\natural, \boldsymbol{Z}^{[3]}_\natural, \boldsymbol{Z}^{[4]}_\natural\}$ of

$$E_{\mathrm{Dir}}(\bar{\boldsymbol{\psi}}^{\mathrm{m}}_\natural) = (0; 0; 0, -\bar{p}_1). \tag{7.4}$$

Since the right-hand side of (7.4) has the parities of a membrane mode (the first component is even and the second odd with respect to x_3), the symmetries of the isotropic elasticity system yield that $\bar{\psi}^{\mathrm{m}}_t$ is even and $\bar{\psi}^{\mathrm{m}}_3$ odd. Thus the asymptotic behavior as $t \to \infty$ has the same parities: only $\boldsymbol{Z}^{[1]}_\natural$ is convenient.

Hence there exists a unique coefficient $c^{\textcircled{1}}_1$ such that $\bar{\boldsymbol{\psi}}^{\mathrm{m}}_\natural$ splits into

$$\bar{\boldsymbol{\psi}}^{\mathrm{m}}_\natural = \bar{\boldsymbol{\varphi}}^{\mathrm{m}}_\natural + c^{\textcircled{1}}_1 \boldsymbol{Z}^{[1]}_\natural \quad \text{with } \bar{\boldsymbol{\varphi}}^{\mathrm{m}}_\natural \text{ exponentially decreasing.} \tag{7.5}$$

Similarly, let $\bar{\boldsymbol{\psi}}^{\mathrm{b}}_\natural$ be the solution in $H^1_\eta(\Sigma^+)^2 \oplus \mathrm{span}\{\boldsymbol{Z}^{[1]}_\natural, \boldsymbol{Z}^{[3]}_\natural, \boldsymbol{Z}^{[4]}_\natural\}$ of

$$E_{\mathrm{Dir}}(\bar{\boldsymbol{\psi}}^{\mathrm{b}}_\natural) = (0; 0; 0, -\bar{p}_2). \tag{7.6}$$

Since the right-hand side of (7.6) has the parities of a bending mode, the symmetries of the problem yield that $\bar{\psi}^{\mathrm{b}}_t$ is odd and $\bar{\psi}^{\mathrm{b}}_3$ even with respect to x_3. Thus only $\boldsymbol{Z}^{[3]}_\natural$ and $\boldsymbol{Z}^{[4]}_\natural$ are present in the asymptotics at infinity of $\bar{\boldsymbol{\psi}}^{\mathrm{b}}_\natural$.

Hence there exist unique coefficients $c^{\textcircled{1}}_3$ and $c^{\textcircled{1}}_4$ such that $\bar{\boldsymbol{\psi}}^{\mathrm{b}}_\natural$ splits into

$$\bar{\boldsymbol{\psi}}^{\mathrm{b}}_\natural = \bar{\boldsymbol{\varphi}}^{\mathrm{b}}_\natural + c^{\textcircled{1}}_3 \boldsymbol{Z}^{[3]}_\natural + c^{\textcircled{1}}_4 \boldsymbol{Z}^{[4]}_\natural \quad \text{with } \bar{\boldsymbol{\varphi}}^{\mathrm{b}}_\natural \text{ exponentially decreasing.} \tag{7.7}$$

Then $\boldsymbol{\psi}^1_\natural$, defined as

$$\boldsymbol{\psi}^1_\natural(t, s, x_3) = \mathrm{div}_*\, \boldsymbol{\zeta}^0_*(s)\ \bar{\boldsymbol{\psi}}^{\mathrm{m}}_\natural(t, x_3) + \Delta_* \zeta^0_3(s)\ \bar{\boldsymbol{\psi}}^{\mathrm{b}}_\natural(t, x_3),$$

is the solution for each $s \in \partial\omega$ of (cf. (7.3), (7.4), and (7.6))

$$E_{\mathrm{Dir}}(\boldsymbol{\psi}^1_\natural) = (0; 0; 0, -v^2_3). \tag{7.8}$$

Thus, if we have for each $s \in \partial\omega$, (cf. (7.5) and (7.7))

$$\begin{pmatrix} \zeta^1_n(s) - x_3 \partial_n \zeta^1_3(s) \\ \zeta^2_3(s) \end{pmatrix} \quad = \quad \mathrm{div}_*\, \boldsymbol{\zeta}^0_*(s)\ c^{\textcircled{1}}_1 \boldsymbol{Z}^{[1]}_\natural\,\big|_{\gamma_0} + \Delta_* \zeta^0_3(s) \left(c^{\textcircled{1}}_3 \boldsymbol{Z}^{[3]}_\natural + c^{\textcircled{1}}_4 \boldsymbol{Z}^{[4]}_\natural \right)\big|_{\gamma_0}\,,$$

i.e.,

$$\begin{pmatrix} \zeta_n^1(s) - x_3\partial_n\zeta_3^1(s) \\ \zeta_3^2(s) \end{pmatrix} = \begin{pmatrix} \operatorname{div}_* \boldsymbol{\zeta}_*^0(s)\, c_1^{\textcircled{1}} - x_3\Delta_*\zeta_3^0(s)\, c_4^{\textcircled{1}} \\ \Delta_*\zeta_3^0(s)\, c_3^{\textcircled{1}} \end{pmatrix}, \tag{7.9}$$

then $\boldsymbol{\varphi}_\natural^1$ defined as

$$\boldsymbol{\varphi}_\natural^1(t,s,x_3) = \operatorname{div}_* \boldsymbol{\zeta}_*^0(s)\, \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{m}}(t,x_3) + \Delta_*\zeta_3^0(s)\, \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}}(t,x_3) \tag{7.10}$$

is the solution of $E_{\mathrm{Dir}}(\boldsymbol{\varphi}_\natural^1(s)) = (0;0;\mathfrak{h}_\natural^1(s))$; see (7.2). Thus we have obtained all the results relating to $\boldsymbol{\zeta}^1$ and $\boldsymbol{\varphi}^1$.

7.2. The nonzero coupling constants. There holds the following lemma.

LEMMA 7.1. *The coefficients $c_1^{\textcircled{1}}$ and $c_4^{\textcircled{1}}$ are nonzero.*

Let us prove first that $c_4^{\textcircled{1}}$ is not zero. Let us denote by $\boldsymbol{Z}_\natural$ the polynomial displacement $\frac{1}{2}\boldsymbol{Z}_\natural^{[7]}$. Thus $\boldsymbol{Z}_\natural$ satisfies

$$E_{\mathrm{Dir}}(\boldsymbol{Z}_\natural) = (0;0;0,\bar{p}_2). \tag{7.11}$$

So, (7.11) joined with (7.6)–(7.7) yields that

$$\boldsymbol{K} := \boldsymbol{Z}_\natural + \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}} + c_3^{\textcircled{1}}\boldsymbol{Z}_\natural^{[3]} + c_4^{\textcircled{1}}\boldsymbol{Z}_\natural^{[4]} \quad \in \quad \ker E_{\mathrm{Dir}}\,.$$

The proof proceeds by computation about the "flux" (cf. (6.6)):

$$\Phi_{t=t_0}(\boldsymbol{u}\,|\,\boldsymbol{v}) := \int_{-1}^{+1} \boldsymbol{T}_\natural^{(0)}(\boldsymbol{u})(t_0,x_3)\cdot\boldsymbol{v}(t_0,x_3)\,dx_3\,. \tag{7.12}$$

We have

$$\boldsymbol{T}_\natural^{(0)}(\boldsymbol{Z}_\natural) = \begin{pmatrix} -4\,\frac{\mu(\lambda+\mu)}{\lambda+2\mu}\,x_3 \\ 0 \end{pmatrix}.$$

Thus,

$$\Phi_{t=0}\Big(\boldsymbol{Z}_\natural\,|\,c_3^{\textcircled{1}}\boldsymbol{Z}_\natural^{[3]} + c_4^{\textcircled{1}}\boldsymbol{Z}_\natural^{[4]}\Big) = \frac{8}{3}\,\frac{\mu(\lambda+\mu)}{\lambda+2\mu}\,c_4^{\textcircled{1}}. \tag{7.13}$$

We are going to prove that (cf. (7.7))

$$\Phi_{t=0}\Big(\boldsymbol{Z}_\natural\,|\,c_3^{\textcircled{1}}\boldsymbol{Z}_\natural^{[3]} + c_4^{\textcircled{1}}\boldsymbol{Z}_\natural^{[4]}\Big) = \Phi_{t=0}\Big(\boldsymbol{K}\,|\,c_3^{\textcircled{1}}\boldsymbol{Z}_\natural^{[3]} + c_4^{\textcircled{1}}\boldsymbol{Z}_\natural^{[4]} + \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}}\Big) \tag{7.14}$$

and that

$$\Phi_{t=0}\Big(\boldsymbol{K}\,|\,c_3^{\textcircled{1}}\boldsymbol{Z}_\natural^{[3]} + c_4^{\textcircled{1}}\boldsymbol{Z}_\natural^{[4]} + \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}}\Big) > 0. \tag{7.15}$$

The fact that $c_4^{\textcircled{1}} > 0$ is clearly a consequence of (7.13)–(7.15).

In order to prove (7.14) and (7.15), we abbreviate the notations by

$$c_3^{\textcircled{1}}\boldsymbol{Z}_\natural^{[3]} + c_4^{\textcircled{1}}\boldsymbol{Z}_\natural^{[4]} := \boldsymbol{R} \quad \text{and} \quad \boldsymbol{\varphi} := \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}}\,.$$

Proof of (7.14). We want to prove that $\Phi_{t=0}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{R}) = \Phi_{t=0}(\boldsymbol{K} \,|\, \boldsymbol{R} + \boldsymbol{\varphi})$. Indeed, integrating by parts on the rectangle $\Sigma_L = (0, L) \times (-1, 1)$ we obtain

$$\begin{aligned}
&\int_{-1}^{+1} \Big[\boldsymbol{T}_\natural^{(0)}(\boldsymbol{K}) \cdot (\boldsymbol{R} + \boldsymbol{\varphi}) - \boldsymbol{K} \cdot \boldsymbol{T}_\natural^{(0)}(\boldsymbol{R} + \boldsymbol{\varphi})\Big]\,(0, x_3)\, dx_3 \\
&\quad - \int_{-1}^{+1} \Big[\boldsymbol{T}_\natural^{(0)}(\boldsymbol{K}) \cdot (\boldsymbol{R} + \boldsymbol{\varphi}) - \boldsymbol{K} \cdot \boldsymbol{T}_\natural^{(0)}(\boldsymbol{R} + \boldsymbol{\varphi})\Big]\,(L, x_3)\, dx_3 \\
&\qquad\qquad = \int_0^L \Big[\mathcal{G}_\natural^{(0)}(\boldsymbol{K}) \cdot (\boldsymbol{R} + \boldsymbol{\varphi}) - \boldsymbol{K} \cdot \mathcal{G}_\natural^{(0)}(\boldsymbol{R} + \boldsymbol{\varphi})\Big]\,(t, 1)\, dt \\
&\qquad\qquad\quad - \int_0^L \Big[\mathcal{G}_\natural^{(0)}(\boldsymbol{K}) \cdot (\boldsymbol{R} + \boldsymbol{\varphi}) - \boldsymbol{K} \cdot \mathcal{G}_\natural^{(0)}(\boldsymbol{R} + \boldsymbol{\varphi})\Big]\,(t, -1)\, dt \\
&\qquad\qquad\qquad\qquad - \int_{\Sigma_L} \mathcal{B}_\natural^{(0)}(\boldsymbol{K}) \cdot (\boldsymbol{R} + \boldsymbol{\varphi}) - \boldsymbol{K} \cdot \mathcal{B}_\natural^{(0)}(\boldsymbol{R} + \boldsymbol{\varphi})\,.
\end{aligned}$$

As $\mathcal{B}_\natural^{(0)}(\boldsymbol{K}) = \mathcal{B}_\natural^{(0)}(\boldsymbol{Z}_\natural) = 0$ and $\mathcal{G}_\natural^{(0)}(\boldsymbol{K}) = \mathcal{G}_\natural^{(0)}(\boldsymbol{Z}_\natural) = 0$, the above right-hand side is zero. Therefore

$$\Phi_{t=0}(\boldsymbol{K} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}) = \Phi_{t=L}(\boldsymbol{K} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}) - \int_{-1}^{+1} \boldsymbol{K}(L, x_3) \cdot \boldsymbol{T}_\natural^{(0)}(\boldsymbol{R} + \boldsymbol{\varphi})(L, x_3)\; dx_3.$$

Since $\boldsymbol{T}_\natural^{(0)}(\boldsymbol{R}) = 0$ ($\boldsymbol{R}$ is a rigid displacement) and since $\boldsymbol{\varphi}$ is exponentially decreasing, we deduce from the identity above that, for all $0 < \eta < \eta_0$

$$\Phi_{t=0}(\boldsymbol{K} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}) = \Phi_{t=L}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{R}) + \mathcal{O}(e^{-\eta L}).$$

But for all L, we have the conservation of the flux against rigid displacements

$$\Phi_{t=L}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{R}) = \Phi_{t=0}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{R}),$$

whence the result. □

Proof of (7.15). We want to prove that $\Phi_{t=0}(\boldsymbol{K} \,|\, \boldsymbol{R}{+}\boldsymbol{\varphi}) > 0$. To see it, notice that, since $\boldsymbol{Z}_\natural\big|_{t=0} = -(\boldsymbol{R}{+}\boldsymbol{\varphi})\big|_{t=0}$ and since we easily check the equality $\Phi_{t=0}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{Z}_\natural) = 0$, we have

$$\begin{aligned}
\Phi_{t=0}(\boldsymbol{K} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}) \;&=\; \Phi_{t=0}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{R} + \boldsymbol{\varphi}) + \Phi_{t=0}(\boldsymbol{R} + \boldsymbol{\varphi} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}), \\
&=\; -\Phi_{t=0}(\boldsymbol{Z}_\natural \,|\, \boldsymbol{Z}_\natural) + \Phi_{t=0}(\boldsymbol{\varphi} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}), \\
&=\; \Phi_{t=L}(\boldsymbol{\varphi} \,|\, \boldsymbol{R} + \boldsymbol{\varphi}) + \int_{\Sigma_L} A\, e(\partial_t, \partial_3)(\boldsymbol{\varphi}) : e(\partial_t, \partial_3)(\boldsymbol{R} + \boldsymbol{\varphi}), \\
&=\; \int_{\Sigma_L} A\, e(\partial_t, \partial_3)(\boldsymbol{\varphi}) : e(\partial_t, \partial_3)(\boldsymbol{\varphi}) + \mathcal{O}(e^{-\eta L}).
\end{aligned}$$

Since $\boldsymbol{Z}_\natural + \boldsymbol{R}$ is clearly not zero on $\{t = 0\}$, then $\boldsymbol{\varphi} \not\equiv 0$. The result follows from the positivity of the elasticity matrix A. □

The positivity of $c_1^{\textcircled{1}}$ can be proved analogously to that of $c_4^{\textcircled{1}}$, taking into account that $\boldsymbol{Z}_\natural^{[5]}$ satisfies problem $E_{\mathrm{Dir}}(\boldsymbol{Z}_\natural^{[5]}) = (0; 0; 0, \bar{p}_1)$, thus

$$\boldsymbol{K}^{\mathrm{m}} := \boldsymbol{Z}_\natural^{[5]} + \bar{\boldsymbol{\varphi}}_\natural^{\mathrm{m}} + c_1^{\textcircled{1}} \boldsymbol{Z}_\natural^{[1]} \quad \in \quad \ker E_{\mathrm{Dir}}$$

and that, moreover, there hold

$$\boldsymbol{T}_\natural^{(0)}\Big(\boldsymbol{Z}_\natural^{[5]}\Big) = \begin{pmatrix} 4\, \frac{\mu(\lambda+\mu)}{\lambda+2\mu} \\ 0 \end{pmatrix} \quad \text{and} \quad \Phi_{t=0}\Big(\boldsymbol{Z}_\natural^{[5]} \,|\, c_1^{\textcircled{1}} \boldsymbol{Z}_\natural^{[1]}\Big) = \frac{8\mu(\lambda + \mu)}{\lambda + 2\mu}\; c_1^{\textcircled{1}}\,.$$

7.3. Soft clamped plates: The first terms in the asymptotics. We now have to take care of the space $\mathcal{R}_{\textcircled{2}}$, which is the space of rigid motions $\boldsymbol{v}$ satisfying the soft clamped plate conditions, i.e., v_n and $v_3 = 0$ on the lateral boundary Γ_0. If the mean surface ω is not a disk or an annulus, $\mathcal{R}_{\textcircled{2}}$ is reduced to $\{0\}$. If ω is a disk or an annulus that we may suppose is centered in 0, $\mathcal{R}_{\textcircled{2}}$ is one-dimensional, generated by the in-plane rotation $(x_2, -x_1, 0)$ and the orthogonality condition (2.11), ensuring uniqueness can be transcribed in Ω into $\int_\Omega \boldsymbol{u}_*(\varepsilon) \cdot (x_2, -x_1)^\top = 0$.

Thus, in this situation, the compatibility conditions on ω for the membrane problems (3.13a) has to be checked and the coherence with the orthogonality condition (2.11) has to be realized by an orthogonality condition for the $\boldsymbol{\zeta}_*^k$ in ω. We refer to [11, section 6] for details.

The behavior of the boundary layer terms is very similar to the hard clamped case because the boundary conditions involving the components $\boldsymbol{\varphi}_\natural$ are Dirichlet as in ①; the only change concerns the lateral component φ_s, which is uncoupled from the previous ones and subject now to lateral Neumann conditions instead of Dirichlet.

7.3.1. The traces of $\boldsymbol{\zeta}^0$. Solving recursively equations (5.13)–(5.14), (5.15), and (5.18), we find first the Dirichlet traces at the order zero: $\zeta_n^0 - x_3\partial_n\zeta_3^0$ and ζ_3^0 are zero on $\partial\omega$. Thus, the Dirichlet conditions concerning $\boldsymbol{\zeta}^0$ are obtained.

The terms T_s^0 and T_s^1 are always zero. Next, condition $T_s^2 = 0$ yields (cf. (5.22b))

$$T_s^{(0)}(\boldsymbol{\varphi}^1) = -T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) + 2\mu x_3 \left(\partial_n + \tfrac{1}{R}\right)\partial_s\zeta_3^0.$$

Taking account of the already known Dirichlet conditions for ζ_3^0, we obtain that φ_s^1 solves the Laplace–Neumann problem on the half-strip:

$$L_{\mathrm{Neu}}(\varphi_s^1) = (0; 0; -T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^0)). \tag{7.16}$$

Since, for each fixed $s \in \partial\omega$, $T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^0)$ is a constant, Proposition 6.6 yields that the only exponentially decreasing solution is $\varphi_s^1 \equiv 0$ obtained with $T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) = 0$ on $\partial\omega$. Then $\boldsymbol{\zeta}^0$ satisfies zero boundary conditions according to Table 2.

7.3.2. The traces of $\boldsymbol{\zeta}^1$. The equations (5.15) for $k = 1$ and for $k = 2$ yield the same condition as in case ① for the trace of ζ_3^1 which must vanish, and the same equations (7.2) linking the couple $\boldsymbol{\varphi}_\natural^1$ and the traces of ζ_n^1, $\partial_n\zeta_3^1$, ζ_3^2. Thus, the result concerning these traces is the same for the hard and soft clamped situations.

As a consequence the coefficients $c_1^{\textcircled{2}}$ and $c_4^{\textcircled{2}}$ are equal to their homologues $c_1^{\textcircled{1}}$ and $c_4^{\textcircled{1}}$ for the hard clamped plate.

Concerning the tangential component, the condition $T_s^3 = 0$ yields (cf. (5.22b))

$$T_s^{(0)}(\boldsymbol{\varphi}^2) = -T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) + 2\mu x_3 \left(\partial_n + \tfrac{1}{R}\right)\partial_s\zeta_3^1 - T_s^{(1)}(\boldsymbol{\varphi}^1).$$

Taking into account the already known trace condition $\zeta_3^1 = 0$, (5.23) leads to the following Neumann problem for the lateral part φ_s^2:

$$L_{\mathrm{Neu}}(\varphi_s^2) = \Big(-(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_s\,; -(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_s\,; - T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) + 2\mu x_3\partial_{sn}\zeta_3^1 - T_s^{(1)}(\boldsymbol{\varphi}^1)\Big). \tag{7.17}$$

Proposition 6.6 yields that φ_s^2 is exponentially decreasing if and only if

$$
\begin{aligned}
T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) = -\frac{1}{2}\Bigg(\int_{\Sigma_+} (\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_s(t,x_3)\,dt\,dx_3 \\
- \int_{\mathbb{R}_+}\Big((\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_s(t,1) - (\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_s(t,-1)\Big)\,dt \\
+ \int_{-1}^{+1} T_s^{(1)}(\boldsymbol{\varphi}^1)(0,x_3) - 2\mu x_3 \partial_{sn}\zeta_3^1(0)\,dx_3\Bigg).
\end{aligned}
\tag{7.18}
$$

Since $\varphi_s^1 = 0$, the terms involved in (7.18) reduce to

$$
(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_s = (\lambda+\mu)\partial_s(\partial_t\varphi_t^1 + \partial_3\varphi_3^1), \quad (\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_s = \mu\,\partial_s\varphi_3^1, \quad T_s^{(1)}(\boldsymbol{\varphi}^1) = \mu\,\partial_s\varphi_t^1.
$$

Since only the *even* terms in x_3 contribute to the integrals in (7.18), we see that we have only to take into consideration the membrane part of $\boldsymbol{\varphi}_\natural^1$, which is equal to $\operatorname{div}_*\boldsymbol{\zeta}_*^0(s)\;\bar{\boldsymbol{\varphi}}_\natural^{\mathrm{m}}(t,x_3)$, cf. (7.10). Thus $T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) = c_2^{\textcircled{2}}\,\partial_s \operatorname{div}_*\boldsymbol{\zeta}_*^0$, with $-\frac{2}{\mu}\,c_2^{\textcircled{2}}$ equal to

$$
\frac{\lambda+\mu}{\mu}\int_{\Sigma_+}(\partial_t\bar\varphi_t^{\mathrm{m}} + \partial_3\bar\varphi_3^{\mathrm{m}})\,dt\,dx_3 - \int_{\mathbb{R}_+}\Big(\bar\varphi_3^{\mathrm{m}}(t,1) - \bar\varphi_3^{\mathrm{m}}(t,-1)\Big)\,dt + \int_{-1}^{+1}\bar\varphi_t^{\mathrm{m}}(0,x_3)\,dx_3.
$$

Formulas of Table 3 concerning case $\textcircled{2}$ are completely proved.

7.3.3. Recursivity. It can be proved like in [8], cf. [11, section 6].

8. Simply supported plates. The space of rigid motions $\mathcal{R}_{\textcircled{3}}$ is reduced to $\{0\}$, whereas $\mathcal{R}_{\textcircled{4}}$ is three-dimensional and spanned by the in-plane rigid motions. Here we only present the analysis for the hard simply supported plate. The main feature of the analysis of the soft simply supported plate is the treatment of compatibility conditions: we refer to [11, section 8] for this.

8.1. Hard simple support: The traces of $\boldsymbol{\zeta}^0$. According to (5.15), $D_3^0 = 0$ yields $\zeta_3^0 = 0$ on $\partial\omega$; then $D_s^0 = 0$ is equivalent to $\zeta_s^0 = 0$ on $\partial\omega$. Next, $D_3^1 = 0$ yields $\zeta_3^1 = 0$ on $\partial\omega$, and $D_s^1 = 0$ provides the equation $L_{\mathrm{Dir}}(\varphi_s^1) = (0;0;-\zeta_s^1)$. Then Proposition 6.4 yields that the only exponentially decreasing solution is $\varphi_s^1 \equiv 0$, obtained with $\zeta_s^1 = 0$ on $\partial\omega$.

Conditions $T_n^2 = 0$ (cf. (5.22b)) and $D_3^2 = 0$ yield that $\boldsymbol{\varphi}_\natural^1$ has to solve

$$
E_{\mathrm{Mix1}}(\boldsymbol{\varphi}_\natural^1) = \Big(0;\;0;\;-T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) + x_3\,M_n(\zeta_3^0),\;-(\zeta_3^2 + v_3^2)\Big).
\tag{8.1}
$$

With formulas (6.8) we can compute the three coefficients δ_3, δ_5, and δ_7, and determine conditions on $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0)$, $M_n(\zeta_3^0)$, and ζ_3^2 so that these three coefficients are zero, ensuring that $\boldsymbol{\varphi}_\natural^1$ is exponentially decaying. We have

$$
\bar\gamma_5\,\delta_5 = \int_{-1}^{+1} -T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) + x_3\,M_n(\zeta_3^0)\,dx_3,
\tag{8.2a}
$$

$$
\bar\gamma_7\,\delta_7 = \int_{-1}^{+1} x_3\,T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) - x_3^2\,M_n(\zeta_3^0)\,dx_3,
\tag{8.2b}
$$

$$
\bar\gamma_3\,\delta_3 = \int_{-1}^{+1} 6\bar p_3(-T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) + x_3\,M_n(\zeta_3^0)) + 6\mu(\bar p_2 + \bar p_3{}')(\zeta_3^2 + v_3^2)\,dx_3.
\tag{8.2c}
$$

With (8.2a) and (8.2b), the conditions $\delta_5 = 0$ and $\delta_7 = 0$ give immediately that $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) = 0$ and $M_n(\zeta_3^0) = 0$ on $\partial\omega$, respectively. Then with the formula $v_3^2 = \bar{p}_1 \operatorname{div}_* \boldsymbol{\zeta}_*^0 + \bar{p}_2 \Delta_* \zeta_3^0$ we can compute from (8.2c)

$$\bar{\gamma}_3\, \delta_3 = -4(\tilde{\lambda} + 2\mu)\Big(\zeta_3^2 - \frac{\tilde{\lambda}}{30\mu}\, \Delta_* \zeta_3^0\Big),$$

whence the relation $30\mu\, \zeta_3^2 = \tilde{\lambda}\, \Delta_* \zeta_3^0$ on $\partial\omega$ ensuring the existence of a unique exponentially decreasing profile solution of (8.1).

But we have on $\partial\omega$

$$T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*) = (\tilde{\lambda} + 2\mu) \operatorname{div}_* \boldsymbol{\zeta}_* + 2\mu(\kappa\, \zeta_n - \partial_s \zeta_s), \tag{8.3a}$$

$$M_n(\zeta_3) = (\tilde{\lambda} + 2\mu)\Delta_* \zeta_3 + 2\mu(\kappa\, \partial_n \zeta_3 - \partial_{ss} \zeta_3). \tag{8.3b}$$

Since ζ_s^0 and ζ_3^0 are zero on $\partial\omega$, then $\partial_s \zeta_s^0$ and $\partial_{ss}\zeta_3^0$ are also zero and since $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) = 0$ and $M_n(\zeta_3^0) = 0$ we deduce from (8.3a) the relations

$$\operatorname{div}_* \boldsymbol{\zeta}_*^0 = -\frac{2\mu}{\tilde{\lambda} + 2\mu}\, \kappa\, \zeta_n^0 \quad \text{and} \quad \Delta_* \zeta_3^0 = -\frac{2\mu}{\tilde{\lambda} + 2\mu}\, \kappa\, \partial_n \zeta_3^0. \tag{8.4}$$

Therefore, with $\bar{\boldsymbol{\varphi}}_\natural^{\mathrm{m}}$ the solution of $E_{\mathrm{Mix1}}(\bar{\boldsymbol{\varphi}}_\natural^{\mathrm{m}}) = \big(0;\ 0;\ 0,\ \frac{2\mu}{\tilde{\lambda}+2\mu}\, \bar{p}_1\big)$, and with $\bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}}$ the solution of $E_{\mathrm{Mix1}}(\bar{\boldsymbol{\varphi}}_\natural^{\mathrm{b}}) = \big(0;\ 0;\ 0,\ \frac{2\mu}{\tilde{\lambda}+2\mu}\, (\frac{\tilde{\lambda}}{30\mu} + \bar{p}_2)\big)$, we obtain the expression in Table 6 of the first boundary layer term.

8.2. The traces of $\boldsymbol{\zeta}^1$. The next relations are deduced from $T_n^3 = 0$ and $D_3^3 = 0$: $\boldsymbol{\varphi}_\natural^2$ has to solve

$$-E_{\mathrm{Mix1}}(\boldsymbol{\varphi}_\natural^2) = \Big(\big(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1\big)_\natural\ ;\ \big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_\natural\ ;\ T_t^{(1)}(\boldsymbol{\varphi}^1) + T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) - x_3\, M_n(\zeta_3^1),\ \zeta_3^3 + v_3^3\Big).$$

Since $\varphi_s^1 = 0$, the terms in the right-hand side reduce to

$$\big(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1\big)_t = -(\lambda + 2\mu)\, \kappa\, \partial_t \varphi_t^1, \quad \big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_t = 0, \quad T_t^{(1)}(\boldsymbol{\varphi}^1) = -\lambda\, \kappa\, \varphi_t^1.$$

The cancellation of the coefficients δ_5, δ_7, and δ_3 (cf. (8.2a)) is ensured by relations determining $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^1)$, $M_n(\zeta_3^1)$, and ζ_3^3. In particular, we have

$$\begin{aligned} T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) = -\frac{1}{2}\Bigg(&\int_{\Sigma_+} \big(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1\big)_t(t, x_3)\, dt\, dx_3 \\ &- \int_{\mathbb{R}_+} \Big(\big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_t(t, 1) - \big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_t(t, -1)\Big)\, dt \\ &+ \int_{-1}^{+1} T_t^{(1)}(\boldsymbol{\varphi}^1)(0, x_3) - x_3 M_n(\zeta_3^1)(0)\, dx_3\Bigg). \end{aligned}$$

Combining with the already known expression for $\boldsymbol{\varphi}^1$, we obtain the formula of Table 3 for $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^1)$. The trace $M_n(\zeta_3^1)$ is determined similarly.

9. Sliding edge. Lateral condition $\textcircled{6}$ is the other one, with $\textcircled{3}$, which allows a reflexion across the boundary in any region $\mathcal{V}$ where it is flat. If the support of the data avoids $\mathcal{V}$, there are *no boundary layer* terms and $\boldsymbol{u}(\varepsilon)$ can be expanded in a power series in $\mathcal{V}$. In the special case when ω is a rectangle (in principle forbidden here!) and the support of the data avoids the lateral boundary, the solution can be extended outside Ω in both in-plane directions into a periodic solution in $\mathbb{R}^2 \times I$: this link is indicated by Paumier in [27], where the periodic boundary conditions are addressed.

If the midplane of the plate ω is not a disk or an annulus, then the space $\mathcal{R}_{⑥}$ is one-dimensional and spanned by the vertical translation $(0,0,1)$. But if ω is a disk or an annulus that we may suppose is centered in 0, then $\mathcal{R}_{⑥}$ is two-dimensional, generated by the vertical translation $(0,0,1)$ and the in-plane rotation $(x_2, -x_1, 0)$. Here we will only treat the generic case.

9.1. The traces of $\boldsymbol{\zeta}^0$. As the Dirichlet trace D_n^0 is zero, we have $\zeta_n^0 = 0$ and $\partial_n \zeta_3^0 = 0$ on $\partial\omega$. We deduce the problem for $\boldsymbol{\varphi}_\natural^1$ from $D_n^1 = 0$ and $T_3^1 = 0$:

$$E_{\text{Mix2}}(\boldsymbol{\varphi}_\natural^1) = (0; 0; -\zeta_n^1 + x_3 \partial_n \zeta_3^1, 0)\,.$$

Proposition 6.11 then yields the conditions $\zeta_n^1 = 0$ and $\partial_n \zeta_3^1 = 0$ on $\partial\omega$ and thus $\boldsymbol{\varphi}_\natural^1 \equiv 0$.

The condition $T_s^2 = 0$ yields that φ_s^1 has to satisfy

$$L_{\text{Neu}}(\varphi_s^1) = (0; 0; -T_s^{\text{m}}(\boldsymbol{\zeta}_*^0) + 2\mu x_3 (\partial_n + \kappa)\partial_s \zeta_3^0)\,. \tag{9.1}$$

Proposition 6.6 yields that $T_s^{\text{m}}(\boldsymbol{\zeta}_*^0) = 0$ on $\partial\omega$. Combining with $\partial_n \zeta_3^0 = 0$ on $\partial\omega$, this solution is given by (cf. Lemma 6.7)

$$\varphi_s^1 = \kappa\, \partial_s \zeta_3^0(s)\, \bar{\varphi}_{\text{Neu}}^{\text{s}}(t, x_3). \tag{9.2}$$

With $T_s^3 = 0$ we obtain that φ_s^2 has to satisfy

$$L_{\text{Neu}}(\varphi_s^2) = \Big(-(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_s\,;\; -(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_s\,;\; \mathfrak{h}_s \Big)\,, \tag{9.3}$$

where the terms in the right-hand side are given by, since $\boldsymbol{\varphi}_\natural^1 = 0$,

$$(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_s = \mu\kappa \Big(\partial_{tt}(t\,\varphi_s^1) + \partial_{33}(t\,\varphi_s^1) - \partial_t \varphi_s^1 \Big)\,, \quad (\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_s = 0,$$
$$\mathfrak{h}_s = -\Big(2\mu\kappa\varphi_s^1 + T_s^{\text{m}}(\boldsymbol{\zeta}_*^1) - 2\mu x_3 (\partial_n + \kappa)\partial_s \zeta_3^1 \Big)\,.$$

With the help of Proposition 6.6 and the fact that φ_s^1 is odd with respect to x_3, we deduce that $T_s^{\text{m}}(\boldsymbol{\zeta}_*^1) = 0$ on $\partial\omega$. Taking into account relation (9.2) and the already known condition $\partial_n \zeta_3^1 = 0$ on $\partial\omega$, this solution is given by

$$\varphi_s^2 = -\kappa^2 \partial_s \zeta_3^0\, \bar{\psi}_{\text{Neu}}^{\text{s}} + \kappa\, \partial_s \zeta_3^1\, \bar{\varphi}_{\text{Neu}}^{\text{s}}\,, \tag{9.4}$$

where $\bar{\psi}_{\text{Neu}}^{\text{s}}$ is the (odd) exponentially decreasing solution of

$$L_{\text{Neu}}(\bar{\psi}_{\text{Neu}}^{\text{s}}) = \mu \Big(\Delta(t\, \bar{\varphi}_{\text{Neu}}^{\text{s}}) - \partial_t \bar{\varphi}_{\text{Neu}}^{\text{s}}\,;\; 0\,;\; 2\bar{\varphi}_{\text{Neu}}^{\text{s}} \Big). \tag{9.5}$$

Conditions $D_n^2 = 0$ and $T_3^2 = 0$ lead to the following problem for $\boldsymbol{\varphi}_\natural^2$:

$$E_{\text{Mix2}}(\boldsymbol{\varphi}_\natural^2) = \Big(-(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_\natural\,;\; -(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_\natural\,;\; \mathfrak{h}_t\,,\; \mathfrak{h}_3 \Big)\,, \tag{9.6}$$

where the terms in the right-hand side are given by

$$\text{(9.7a)}\qquad (\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_t = (\lambda+\mu)\,\partial_t\partial_s\varphi^1_s\,, \qquad (\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_t = 0,$$

$$\text{(9.7b)}\qquad (\mathcal{B}^{(1)}\boldsymbol{\varphi}^1)_3 = (\lambda+\mu)\,\partial_3\partial_s\varphi^1_s, \qquad (\mathcal{G}^{(1)}\boldsymbol{\varphi}^1)_3 = \lambda\,\partial_s\varphi^1_s,$$

$$\text{(9.7c)}\qquad \mathfrak{h}_t = -\left(\zeta^2_n - x_3\partial_n\zeta^2_3 + \bar p_2\,\partial_n\operatorname{div}_*\boldsymbol{\zeta}^0_* + \bar p_3\,\partial_n\Delta_*\zeta^0_3 + (G(\boldsymbol{f},\boldsymbol{g}^\pm))_n\right),$$

$$\text{(9.7d)}\qquad \mathfrak{h}_3 = -\mu\left((\bar p_2 + \bar p_3')\,\partial_n\Delta_*\zeta^0_3 + \partial_3(G(\boldsymbol{f},\boldsymbol{g}^\pm))_n\right)\,.$$

Combining with (9.2), the condition $\delta_8 = 0$ from Proposition 6.11 yields

$$2\mu\partial_s(\partial_n+\kappa)\partial_s\zeta^0_3\int_{\mathbb{R}^+}\bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}(t,1)\,dt = \int_{-1}^{+1}\mathfrak{h}_3\,dx_3\,.$$

Using the expressions of G_n (cf. Definition 4.5) and of $\bar p_2$ and $\bar p_3$ (cf. (4.4)), we derive

$$\int_{-1}^{+1}\mathfrak{h}_3\,dx_3 = -\left[-\frac{2}{3}(\tilde\lambda+2\mu)\partial_n\Delta_*\,\zeta^0_3 + \int_{-1}^{+1}x_3\,f_n\,dx_3 + g^+_n + g^-_n\right]\Bigg|_{\partial\omega}.$$

Then Lemma 6.7 yields

$$\frac{2}{3}\underbrace{\left((\tilde\lambda+2\mu)\partial_n\Delta_*\,\zeta^0_3 + 2\mu\partial_s(\partial_n+\kappa)\partial_s\zeta^0_3\right)}_{=\,N_n(\zeta^0_3)} = \left(\int_{-1}^{+1}x_3\,f_n\,dx_3 + g^+_n + g^-_n\right)\Bigg|_{\partial\omega},$$

hence the condition $N_n(\zeta^0_3) = \frac{3}{2}\int_{-1}^{+1}x_3\,f_n\,dx_3 + g^+_n + g^-_n$ on $\partial\omega$. Then the compatibility condition for the solvability of problem (3.13b) for ζ^0_3 reads

$$\text{(9.8)}\qquad \int_\omega R^0_{\mathrm{b}}(x_*)\,dx_* - \int_{\partial\omega}\frac{3}{2}\Big(\int_{-1}^{+1}x_3\,f_n\,dx_3 + g^+_n + g^-_n\Big)(0,s)\,ds = 0\,.$$

With the help of the divergence theorem and formula (4.9), we can rewrite (9.8) as

$$\frac{3}{2}\int_\omega\left\{\int_{-1}^{+1}f_3\,dx_3 + g^+_3 - g^-_3\right\}dx_* = 0\,,$$

which is nothing else than the compatibility condition (2.10), whence (9.8).

9.2. The traces of $\boldsymbol{\zeta}^1$. The only remaining boundary condition is that for $N_n(\zeta^1_3)$. Therefore we only consider the problem for $\boldsymbol{\varphi}^3_\natural$, which is deduced from $D^3_n = 0$ and $T^3_3 = 0$ and reads

$$E_{\mathrm{Mix2}}(\boldsymbol{\varphi}^3_\natural) = \Big(-(\mathcal{B}^{(1)}\boldsymbol{\varphi}^2)_\natural - (\mathcal{B}^{(2)}\boldsymbol{\varphi}^1)_\natural\,;\; -(\mathcal{G}^{(1)}\boldsymbol{\varphi}^2)_\natural - (\mathcal{G}^{(2)}\boldsymbol{\varphi}^1)_\natural\,;\; \mathfrak{h}_t\,,\;\mathfrak{h}_3\Big)\,.$$

The boundary condition prescribing $N_n(\zeta^1_3)$ is then found by the cancellation of the coefficient δ_8 (6.9a). For this, we need an expression for $\boldsymbol{\varphi}^2_\natural$, which is derived from the cancellation of the constants δ_1 and δ_4 (6.9b)–(6.9c) relating to problem (9.6). The details can be found in [12, section 5].

Let us check the compatibility condition for ζ^1_3. Setting $\boldsymbol{\varphi} = \boldsymbol{\varphi}^1 + \varepsilon\boldsymbol{\varphi}^2$, we have by construction

$$\text{(9.9)}\qquad \begin{aligned} N_n(\zeta^0_3+\varepsilon\zeta^1_3) = \frac{3}{2}\Big(&\int_{\Sigma^+}\mathfrak{f}_3(\varepsilon) - \int_{\mathbb{R}^+}\Big(\mathfrak{g}^+_3(\varepsilon) - \mathfrak{g}^-_3(\varepsilon)\Big) + \int_{-1}^{+1}\mathfrak{h}_3(\varepsilon)\Big) \\ &+ 2\mu\,\partial_s(\partial_n+\kappa)\partial_s(\zeta^0_3+\varepsilon\zeta^1_3), \end{aligned}$$

TABLE 9
Auxiliary problems.

⑤	$\operatorname{mes}(\omega)\, L^{\mathrm{b}}(\eta_\omega) = 1 \quad \text{in } \omega$	$\eta_\omega = 0 \;\text{ and }\; \partial_n \eta_\omega = 0 \quad \text{on } \partial\omega$
⑦	$\operatorname{mes}(\omega)\, L^{\mathrm{b}}(\xi_\omega) = 1 \quad \text{in } \omega$	$\xi_\omega = 0 \;\text{ and }\; M_n(\xi_\omega) = 0 \quad \text{on } \partial\omega$

TABLE 10
Boundary conditions.

⑤	$\zeta_3^1 = c_3^{\textcircled{5}} \Big(\oint_{\partial\omega} \oint_{\partial\omega} L - \int_{\partial\omega} \Big(\oint_{\partial\omega} \oint_{\partial\omega} L \Big) N_n(\eta_\omega) \Big)$	$\partial_n \zeta_3^1 = 0$
⑦	$\zeta_3^1 = c_3^{\textcircled{7}} \Big(\oint_{\partial\omega} \oint_{\partial\omega} L + 2\mu \int_{\partial\omega} L\, \partial_n \xi_\omega \; - \int_{\partial\omega} \Big(\oint_{\partial\omega} \oint_{\partial\omega} L \Big) N_n(\xi_\omega) \Big)$	$M_n(\zeta_3^1) = c_4^{\textcircled{7}}\, L$

where

$$\mathfrak{f}(\varepsilon) = \mathcal{B}\boldsymbol{\varphi} + \mathcal{O}(\varepsilon^2), \qquad \mathfrak{g}(\varepsilon) = \mathcal{G}\boldsymbol{\varphi} + \mathcal{O}(\varepsilon^2), \qquad \mathfrak{h}(\varepsilon) = \boldsymbol{T}\boldsymbol{\varphi} + \mathcal{O}(\varepsilon^2).$$

With $\boldsymbol{w}(\tilde{x}) = \chi(r)\, \boldsymbol{\varphi}(\frac{r}{\varepsilon}, s, \frac{\tilde{x}_3}{\varepsilon})$ on Ω^ε and integrating (9.9) along $\partial\omega$, we obtain for any rigid motion $\boldsymbol{v} = (0, 0, a)$ in $\mathcal{R}_{\textcircled{6}}$

$$\int_{\partial\omega} N_n(\zeta_3^0 + \varepsilon\zeta_3^1)\, v_3 = -\frac{3}{2} \int_{\Omega^\varepsilon} Ae(\boldsymbol{w}) : e(\boldsymbol{v}) + \mathcal{O}(\varepsilon^2) = \mathcal{O}(\varepsilon^2),$$

where we have used $\int_{\partial\omega} \partial_s(\partial_n + \kappa)\partial_s(\zeta_3^0 + \varepsilon\zeta_3^1)\, ds = 0$. The desired compatibility condition then follows.

10. Friction conditions. We only give a few precisions about the traces of the first Kirchhoff–Love generators $\boldsymbol{\zeta}^0$ and $\boldsymbol{\zeta}^1$ for conditions ⑤ and ⑦, referring to [12, sections 4 and 6] for the proofs, which make use in particular of Lemma 6.5.

The membrane boundary operators $\gamma^{\mathrm{m},j}$, $j = 1, 2$, are Dirichlet's in both cases and the corresponding traces $\gamma^0_{\mathrm{m},j}$ and $\gamma^1_{\mathrm{m},j}$ are zero.

The spaces of rigid motions $\mathcal{R}_{\textcircled{5}}$ and $\mathcal{R}_{\textcircled{7}}$ are one-dimensional and both are generated by the vertical translation $(0, 0, 1)$. As a consequence, the first terms ζ_3^0 and ζ_3^1 have to satisfy the zero mean value condition on ω. The bending boundary operators $\gamma^{\mathrm{b},j}$, $j = 1, 2$, are Dirichlet's for ⑤, and the trace operator on $\partial\omega$ and M_n for ⑦. Thus the corresponding problems (3.13b) are uniquely solvable. The way out is that the boundary conditions issued from the solution of the Ansatz include $\partial_s \zeta_3 = 0$ on $\partial\omega$. Thus the trace of ζ_3 can be fixed to any constant (we assume here for simplicity that $\partial\omega$ is connected), which can be chosen such that $\int_\omega \zeta_3 = 0$. The formula for this constant relies on the introduction of the solutions η_ω and ξ_ω of the auxiliary problems (Table 9).

NOTATION 10.1. *If L is an integrable function on $\partial\omega$ such that $\int_{\partial\omega} L = 0$, then we denote by $\oint_{\partial\omega} L$ the unique primitive of L along $\partial\omega$ with zero mean value on $\partial\omega$ (that is, $\int_{\partial\omega} \oint L\, ds = 0$). The second primitive $\oint_{\partial\omega} \oint_{\partial\omega} L$ then makes sense.*

For condition ⑤, $\partial_n \zeta_3^0 = 0$ and ζ_3^0 is equal to the constant $-\int_\omega R_{\mathrm{b}}^0 \eta_\omega$ on $\partial\omega$, whereas for condition ⑦, $M_n(\zeta_3^0) = 0$ and ζ_3^0 is equal to the constant $-\int_\omega R_{\mathrm{b}}^0\, \xi_\omega$ on $\partial\omega$. Finally, the boundary conditions for ζ_3^1 are displayed in Table 10, with L given in (3.15).

11. Free. The space $\mathcal{R}_{⑧}$ is six-dimensional and spanned by all rigid motions. We are only going to explain how the traces of $\boldsymbol{\zeta}^0$ can be determined by our method and refer to [12, section 7] for the traces of $\boldsymbol{\zeta}^1$. The nonhomogeneity of the boundary condition $N_n(\zeta_3^0)$ is known; see Ciarlet [4, Theorem 1.7.2].

From the conditions $T_3^1 = 0$ and $T_n^2 = 0$, we obtain for $\boldsymbol{\varphi}_\natural^1$

$$E_{\mathrm{Free}}(\boldsymbol{\varphi}_\natural^1) = (0; 0; -T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) + x_3 M_n(\zeta_3^0), 0)\,. \tag{11.1}$$

From the cancellation of the constants δ_5 and δ_7 in Proposition 6.12, the conditions $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) = 0$ and $M_n(\zeta_3^0) = 0$ on $\partial\omega$ are obtained. Thus $\boldsymbol{\varphi}_\natural^1 \equiv 0$.

The condition $T_s^2 = 0$ yields that φ_s^1 has to satisfy problem (9.1). Thus $T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^0) = 0$ on $\partial\omega$ and φ_s^1 is then given by (cf. Lemma 6.7)

$$\varphi_s^1 = (\partial_n + \kappa)\partial_s\zeta_3^0(s)\,\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}\,. \tag{11.2}$$

With $T_s^3 = 0$ we obtain that φ_s^2 has to satisfy problem (9.3), hence the condition $T_s^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) = 0$ on $\partial\omega$ ensures the existence of an exponentially decaying profile. Taking into account the relation (11.2), this solution is given by

$$\varphi_s^2 = -\kappa(\partial_n + \kappa)\partial_s\zeta_3^0\,\bar\psi_{\mathrm{Neu}}^{\mathrm{s}} + (\partial_n + \kappa)\partial_s\zeta_3^1\,\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}\,, \tag{11.3}$$

where $\bar\psi_{\mathrm{Neu}}^{\mathrm{s}}$ is the solution of problem (9.5).

The conditions $T_3^2 = 0$ and $T_n^3 = 0$ lead to the following problem for $\boldsymbol{\varphi}_\natural^2$:

$$E_{\mathrm{Free}}(\boldsymbol{\varphi}_\natural^2) = \Big(-\big(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1\big)_\natural\,;\ -\big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_\natural\,;\ \mathfrak{h}_t\,,\ \mathfrak{h}_3\Big)\,, \tag{11.4}$$

where the terms in the right-hand side of (11.4) are given by

$$\big(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1\big)_t = (\lambda+\mu)\,\partial_t\partial_s\varphi_s^1\,,\ \big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_t = 0\,,\ \mathfrak{h}_t = -\big(\lambda\partial_s\varphi_s^1 + T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) - x_3 M_n(\zeta_3^1)\big)\,,$$

whereas $\big(\mathcal{B}^{(1)}\boldsymbol{\varphi}^1\big)_3$ and $\big(\mathcal{G}^{(1)}\boldsymbol{\varphi}^1\big)_3$ are still given by (9.7b) and $\mathfrak{h}_3$ by (9.7d). Thus, the cancellation of the constants δ_5, δ_7, and δ_8 from Proposition 6.12 is required. The cancellation of δ_5 leads to the boundary condition $T_n^{\mathrm{m}}(\boldsymbol{\zeta}_*^1) = 0$ on $\partial\omega$. Inserting the expressions involved, the condition $\delta_7 = 0$ reads

$$\begin{aligned}\Bigg[(\lambda+\mu)\int_{\Sigma^+}(x_3\,\partial_t\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}} - t\,\partial_3\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}})\,dt\,dx_3 + \lambda\int_0^\infty t\,(\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(1,t) - \bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(1,t))\,dt\\ + \lambda\int_{-1}^{+1} x_3\,\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(0,x_3)\,dx_3\Bigg]\,\partial_s(\partial_n+\kappa)\partial_s\zeta_3^0 - \int_{-1}^{+1} x_3^2\,M_n(\zeta_3^1)\,dx_3 = 0\,.\end{aligned}$$

As the boundary layer term $\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}$ is odd, the above condition becomes

$$\frac{2}{3}\,M_n(\zeta_3^1) = \partial_s(\partial_n+\kappa)\partial_s\zeta_3^0\left[-\mu\int_{-1}^{+1} x_3\,\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(0,x_3)\,dx_3 - 2\mu\int_0^\infty t\ \bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(1,t)\,dt\right].$$

Applying the second Green formula for Laplace to the functions $\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(t,x_3)$ and $w(t,x_3) = t\,x_3$ yields the relation

$$2\int_0^\infty t\,\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(t,1)\,dt = \int_{-1}^{+1} x_3\,\bar\varphi_{\mathrm{Neu}}^{\mathrm{s}}(0,x_3)\,dx_3\,.$$

Thus $M_n(\zeta_3^1) = c_3^{\textcircled{8}} \partial_s(\partial_n + \kappa)\partial_s \zeta_3^0$ on $\partial\omega$ with $c_3^{\textcircled{8}} = -3\mu \int_{-1}^{+1} x_3\, \bar\varphi^{\mathrm{s}}_{\mathrm{Neu}}(0, x_3)\, dx_3$.

The evaluation of the condition $\delta_8 = 0$ has been already done in section 8.1, which yields in exactly the same way, formula (3.14) for the trace $N_n(\zeta_3^0)$.

Now let us check the compatibility conditions ensuring the existence of the generator $\boldsymbol{\zeta}^0$. Concerning $\boldsymbol{\zeta}_*^0$, we have to show that the membrane right-hand side $\boldsymbol{R}_{\mathrm{m}}^0$ of the limit problem is orthogonal to each of the two-dimensional rigid motions $(1,0)$, $(0,1)$, and $(x_2, -x_1)$, since we have homogeneous traction boundary conditions in the problem for $\boldsymbol{\zeta}_*^0$. These orthogonality conditions are clearly a consequence of the expression of the right-hand side $\boldsymbol{R}_{\mathrm{m}}^0$ and of the three-dimensional compatibility conditions (2.10) for in-plane rigid motions.

The compatibility conditions for ζ_3^0 remains to be checked. They are related to the kernel of L^{b} with boundary conditions M_n and N_n, i.e., to the functions 1, x_1, and x_2. It has been already shown in section 8.1 that the condition (9.8) relating to the element 1 of the kernel is fulfilled. Now let us check the condition for x_1, namely,

$$\int_\omega x_1\, R_{\mathrm{b}}^0(x_*)\, dx_* - \frac{3}{2} \int_{\partial\omega} x_1 \left(\int_{-1}^{+1} x_3\, f_n\, dx_3 + g_n^+ + g_n^- \right)(0, s)\, ds = 0\ .$$

With the help of the divergence theorem we can rewrite it as

$$\frac{3}{2} \left\{ \int_\Omega (x_1\, f_3 - x_3\, f_1)\, dx_3\, dx_* + \int_\omega \left\{ x_1\, (g_3^+ - g_3^-) - (g_1^+ + g_1^-) \right\}\, dx_* \right\} = 0\,,$$

which coincides with a compatibility condition (2.10) for the three-dimensional problem. Of course, the condition for x_2 can be proved analogously.

12. Error estimates. We provide in this section estimates in H^1 and L^2 norms.

12.1. In $\boldsymbol{H^1}$ norm. In this section we prove Theorem 3.2, which yields an optimal estimation of the error between the scaled displacement $\boldsymbol{u}(\varepsilon)$ and the Ansatz of order N. This extends the results obtained in [8, section 5] for the hard clamped situation to the eight "canonical" boundary conditions on the lateral side. The proof relies on energy estimates and on a very simple argument consisting in pushing the development a few terms further.

We define the space $\mathcal{V}_{\textcircled{i}}(\Omega)$ as the subspace of the admissible displacements $\boldsymbol{u}$ in $V_{\textcircled{i}}(\Omega)$ which are orthogonal for the L^2 product to all the rigid motions $\boldsymbol{v} \in \mathcal{R}_{\textcircled{i}}(\Omega)$. Thus $\boldsymbol{u}(\varepsilon)$ belongs to $\mathcal{V}_{\textcircled{i}}(\Omega)$. Combining Korn's inequality without boundary conditions and the infinitesimal rigid displacement lemma we obtain a Korn inequality with boundary conditions for arbitrary $\boldsymbol{u} \in \mathcal{V}_{\textcircled{i}}(\Omega)$; compare [25] and [4], which reads in terms of the scaled linearized strain tensor $\theta(\varepsilon)$,

$$\left(\int_\Omega A\theta(\varepsilon)(\boldsymbol{u}) : \theta(\varepsilon)(\boldsymbol{u}) \right)^{1/2} \geq C^* \|\theta(\varepsilon)(\boldsymbol{u})\|_{L^2(\Omega)^9} \geq C \|\boldsymbol{u}\|_{H^1(\Omega)}\,. \tag{12.1}$$

Defining the remainder at the order N of the asymptotics of $\boldsymbol{u}(\varepsilon)$ by $\overline{U}^N(\varepsilon) := \boldsymbol{u}(\varepsilon) - U^N(\varepsilon)$, where $U^N(\varepsilon)$ denotes the asymptotic expansion of order N, namely,

$$U^N(\varepsilon) \;=\; \underbrace{\sum_{k=0}^{N} \varepsilon^k\, \underline{\boldsymbol{u}}^k}_{=:\, V^N(\varepsilon)} \;+\; \chi(r) \underbrace{\sum_{k=1}^{N} \varepsilon^k\, \boldsymbol{w}^k \left(\frac{r}{\varepsilon}, s, x_3 \right)}_{=:\, W^N(\varepsilon)} \tag{12.2}$$

with $\underline{\boldsymbol{u}}^k := \boldsymbol{u}^k_{\mathrm{KL}} + \boldsymbol{v}^k$; compare section 3.1 for notations, we only need to establish an a priori estimate for $\overline{U}^N(\varepsilon)$ in the norm of the space $H^1(\Omega)^3$.

Therefore, we split $U^N(\varepsilon)$ into its natural parts $U^N(\varepsilon) = V^N(\varepsilon) + \chi(r)\, W^N(\varepsilon)$. Considering carefully the construction algorithm, in particular the derivation of the boundary layer terms, we observe that for any $N \in \mathbb{N}$, $U^N(\varepsilon)$ belongs to the space $\mathcal{V}_{\textcircled{i}}(\Omega)$. Thus, we have

$$\forall N \in \mathbb{N}\,, \quad \overline{U}^N(\varepsilon) \in \mathcal{V}_{\textcircled{i}}(\Omega)$$

and the variational form of the problem for $\overline{U}^N(\varepsilon)$ can be written down, where we split the deviation to the true solution into an error generated by $V^N(\varepsilon)$ and an error coming from $W^N(\varepsilon)$; compare [8, (5.8)–(5.11)]. For the choice $\boldsymbol{v} = \overline{U}^N(\varepsilon)$ of the test function in the variational formulation of the problem for $\overline{U}^N(\varepsilon)$, we obtain as one side of the resulting equation the energy associated to the remainder, namely,

$$\int_\Omega A\,\theta(\varepsilon)(\overline{U}^N(\varepsilon)) \,:\, \theta(\varepsilon)(\overline{U}^N(\varepsilon))\,.$$

Korn's inequality (12.1) and the coercivity of the operator of elasticity then provides the following rough estimate

$$\|\overline{U}^N(\varepsilon)\|_{H^1(\Omega)^3} \le C\varepsilon^{N-3}$$

exactly in the same manner as in the proof of Lemma 5.3 in [8]. This estimate reads for $\|\overline{U}^{N+4}(\varepsilon)\|_{H^1(\Omega)^3} \le C\varepsilon^{N+1}$ at the rank $N+4$, whence

$$\begin{aligned}(12.3)\quad &\left\|\boldsymbol{u}(\varepsilon)(x) - \boldsymbol{u}^0_{\mathrm{KL}}(x) - \sum_{k=1}^{N} \varepsilon^k \boldsymbol{u}^k\left(x, \frac{r}{\varepsilon}\right)\right\| H^1(\Omega)^3 \\ &\qquad \le C\,\varepsilon^{N+1} + \sum_{k=N+1}^{N+4} \varepsilon^k \left(\|\underline{\boldsymbol{u}}^k\| H^1(\Omega)^3 + \left\|\chi(r)\boldsymbol{w}^k\left(\frac{r}{\varepsilon}, s, x_3\right)\right\| H^1(\Omega)^3 \right)\,.\end{aligned}$$

With the help of the following H^1-estimates of each term in the asymptotics

$$(12.4)\qquad \|\underline{\boldsymbol{u}}^k\|_{H^1(\Omega)^3} \le C \quad\text{and}\quad \left\|\chi(r)\boldsymbol{w}^k\left(\frac{r}{\varepsilon}, s, x_3\right)\right\| H^1(\Omega)^3 \le C\varepsilon^{-1/2}\,,$$

the estimate (3.4) directly follows from (12.3).

12.2. In other norms. The L^2-estimates of each term corresponding to (12.4)

$$(12.5)\qquad \|\underline{\boldsymbol{u}}^k\|_{L^2(\Omega)^3} \le C \quad\text{and}\quad \left\|\chi(r)\boldsymbol{w}^k\left(\frac{r}{\varepsilon}, s, x_3\right)\right\| L^2(\Omega)^3 \le C\varepsilon^{1/2}$$

lead in a straightforward way to the following estimates in L^2 norm:

$$(12.6)\qquad \left\|\boldsymbol{u}(\varepsilon) - \sum_{k=0}^{N} \varepsilon^k\, \underline{\boldsymbol{u}}^k - \chi(r) \sum_{k=1}^{N} \varepsilon^k\, \boldsymbol{w}^k\left(\frac{r}{\varepsilon}, s, x_3\right)\right\| L^2(\Omega)^3 \le C\,\varepsilon^{N+1}\,.$$

The question of estimates in higher norms, H^2 for instance, is also considered in [9] for the clamped case. Such estimates require a splitting of the solution and of terms in the asymptotics, since in general the H^2 regularity is not attained. The situation is similar for all lateral conditions. Let us just emphasize that all the terms in the outer expansion are smooth, but also that the singularities along the edges $\partial\omega \times \{\pm 1\}$ of the plate are concentrated in the inner expansion: the model profiles are all nonsmooth, with a regularity between $H^{3/2}$ and H^3. For example, $\bar{\varphi}^{\rm s}_{\rm Dir}$ is almost H^2 and $\bar{\varphi}^{\rm s}_{\rm Neu}$ is almost H^3, whereas the profiles $\bar{\boldsymbol{\varphi}}^{\rm m}_{{\rm Dir},\natural}$ and $\bar{\boldsymbol{\varphi}}^{\rm b}_{{\rm Dir},\natural}$ occurring in the clamped plates have less regularity, cf. [10].

13. Conclusions. Coming back to the family of thin domains Ω^ε, we will briefly address the question of the determination of a limit solution and of the evaluation of the relative error between this limit and the three-dimensional solution. The correct answer depends on the norm in which the error is evaluated and on the type of the loading.

13.1. $\boldsymbol{H^1}$ norm. We have first to evaluate the behavior of the $H^1(\Omega^\varepsilon)$ norm denoted by $\|\cdot\|_{H1}$ of each of the four types of components of series (3.5), namely, $\tilde{\boldsymbol{u}}^k_{\rm KL,b}$, $\tilde{\boldsymbol{u}}^k_{\rm KL,m}$, $\tilde{\boldsymbol{v}}^k$, and $\boldsymbol{\varphi}^k$. We find

$$\|\tilde{\boldsymbol{u}}^k_{\rm KL,b}\|_{H1} = \mathcal{O}(\varepsilon^{-1/2}), \qquad \|\tilde{\boldsymbol{u}}^k_{\rm KL,m}\|_{H1} = \mathcal{O}(\varepsilon^{1/2}),$$

$$\|\tilde{\boldsymbol{v}}^k\|_{H1} = \mathcal{O}(\varepsilon^{-1/2}), \qquad \|\boldsymbol{\varphi}^k\|_{H1} = \mathcal{O}(1).$$

In the case of a bending load such that $R^0_{\rm b}$ (cf. (4.9)) is nonzero, we have

$$\frac{\|\boldsymbol{u}^\varepsilon - \tilde{\boldsymbol{u}}^0_{\rm KL,b}\|_{H1}}{\|\boldsymbol{u}^\varepsilon\|_{H1}} \le C\,\varepsilon \tag{13.1}$$

and this estimate is sharp for any lateral boundary condition, since the main contribution to the error comes from $\tilde{\boldsymbol{v}}^1$, which is equal to $(0, 0, \bar{p}_2(x_3)\,\Delta_*\zeta^0_3)$; indeed, since we assumed that $R^0_{\rm b}$ is nonzero, $\Delta^2_*\zeta^0_3$ is nonzero, and $\tilde{\boldsymbol{v}}^1 \not\equiv 0$.

In the case of a membrane load such that $\boldsymbol{R}^0_{\rm m}$ (cf. (4.6)) is nonzero, we have to include $\tilde{\boldsymbol{v}}^1$ in the limit solution to have a convergence: we set

$$\boldsymbol{u}^{\rm lim}_{\rm m} = \tilde{\boldsymbol{u}}^0_{\rm KL,m} + \varepsilon\tilde{\boldsymbol{v}}^1 = (\zeta^0_*,\ \bar{p}_1(\tilde{x}_3)\,{\rm div}_*\,\zeta^0_*). \tag{13.2}$$

Then

$$\frac{\|\boldsymbol{u}^\varepsilon - \boldsymbol{u}^{\rm lim}_{\rm m}\|_{H1}}{\|\boldsymbol{u}^\varepsilon\|_{H1}} \le C\,\varepsilon^{1/2}, \quad \text{in cases ①–④}, \tag{13.3}$$

this estimate being generically optimal, in the sense that it is sharp when $\boldsymbol{\varphi}^1$ is nonzero, i.e., when ${\rm div}_*\,\zeta^0_*$ is nonzero on $\partial\omega$ in cases ①, ②, and ④, and when $\kappa\zeta^0_n$ is nonzero on $\partial\omega$ in cases ③. On the other hand,

$$\frac{\|\boldsymbol{u}^\varepsilon - \boldsymbol{u}^{\rm lim}_{\rm m}\|_{H1}}{\|\boldsymbol{u}^\varepsilon\|_{H1}} \le C\,\varepsilon, \quad \text{in cases ⑤–⑧}, \tag{13.4}$$

this estimate being generically optimal too, in the sense that it is sharp when $\tilde{\boldsymbol{v}}^2$ is nonzero, i.e., when ${\rm div}_*\,\zeta^0_* \not\equiv 0$; compare also with [22] for a special membrane loading on a free plate.

13.2. Energy norm. We now set $\|\boldsymbol{u}\|_E = \left(\int_{\Omega^\varepsilon} Ae(\boldsymbol{u}) : e(\boldsymbol{u})\right)^{1/2}$. The energy of the four types of terms in the series (3.5) has the same behavior as their H^1 norm *except* the one concerning $\tilde{\boldsymbol{u}}^k_{\mathrm{KL,b}}$, whose energy is one order smaller:

$$\|\tilde{\boldsymbol{u}}^k_{\mathrm{KL,b}}\|_E = \mathcal{O}(\varepsilon^{1/2}).$$

We obtain exactly the same conclusions if we use this energy, or the L^2 norm of the strain tensor, or the complementary energy. We have to include the polynomial terms up to the order 2 to obtain a convergence: we set $\boldsymbol{u}^{\mathrm{lim}}_{\mathrm{m}}$ as above in (13.2) and moreover

$$\boldsymbol{u}^{\mathrm{lim}}_{\mathrm{b}} = \tilde{\boldsymbol{u}}^0_{\mathrm{KL,b}} + \varepsilon\tilde{\boldsymbol{v}}^1 = (-x_3\nabla_*\zeta^0_3,\ \varepsilon^{-1}\zeta^0_3 + \varepsilon\bar{p}_2(x_3)\,\Delta_*\zeta^0_3); \tag{13.5}$$

cf. [28] and [30] in this context.

In the case of a bending load such that R^0_{b} is nonzero, we have

$$\frac{\|\boldsymbol{u}^\varepsilon - \boldsymbol{u}^{\mathrm{lim}}_{\mathrm{b}}\|_E}{\|\boldsymbol{u}^\varepsilon\|_E} \le C\,\varepsilon^{1/2}, \tag{13.6}$$

this estimate being generically optimal, in the sense that it is sharp when $\boldsymbol{\varphi}^1$ is nonzero, i.e., when ℓ^{b} is nonzero on $\partial\omega$ in cases ①–④ (cf. Table 6) and when ℓ^{s} is nonzero on $\partial\omega$ in cases ⑤–⑧ (cf. Table 7).

In the case of a membrane load such that $\boldsymbol{R}^0_{\mathrm{m}}$ is nonzero, we have exactly the same behavior as with the H^1 norm; see (13.3) and (13.4). In particular, the condition for the optimality of the estimates is visibly sharp, which brings a conclusion to the work [2].

The observation of the first terms in the asymptotics also sheds light on the order of magnitude of the answer of the plate under the loading. The maximal answer rate (of order ε^{-2}) is obtained with a bending load such that R^0_{b} is nonzero and corresponds to the flexural nature of plates. In contrast, the membrane (or stretching) answer is of order 1 when $\boldsymbol{R}^0_{\mathrm{m}}$ is nonzero. Moreover, there are very many other types of loading (bending or membrane) whose answer rate is much lower; see [6].

REFERENCES

[1] D. N. Arnold and R. S. Falk, *Asymptotic analysis of the boundary layer for the Reissner-Mindlin plate model.*, SIAM J. Math. Anal., 27 (1996), pp. 486–514.

[2] C. Chen, *Asymptotic convergence rates for the Kirchhoff plate model*, Ph.D. thesis, Pennsylvania State University, State College, PA, 1995.

[3] P. G. Ciarlet, *Plates and Junctions in Elastic Multi-Structures: An Asymptotic Analysis*, R.M.A. Vol. 14, Masson and Springer-Verlag, Paris, Heidelberg, 1990.

[4] P. G. Ciarlet, *Mathematical Elasticity. Vol.* II, *Theory of Plates*, North-Holland, Amsterdam, 1997.

[5] P. G. Ciarlet and P. Destuynder, *A justification of the two-dimensional plate model*, J. Mécanique, 18 (1979), pp. 315–344.

[6] M. Dauge, I. Djurdjevic, and A. Rössle, *Higher order bending and membrane responses of thin linearly elastic plates*, C. R. Acad. Sci. Paris, Sér. I, 326 (1998), pp. 519–524.

[7] M. Dauge and I. Gruais, *Développement asymptotique d'ordre arbitraire pour une plaque élastique mince encastrée*, C. R. Acad. Sci. Paris, Sér. I, 321 (1995), pp. 375–380.

[8] M. Dauge and I. Gruais, *Asymptotics of arbitrary order for a thin elastic clamped plate.* I: *Optimal error estimates*, Asymptot. Anal., 13 (1996), pp. 167–197.

[9] M. Dauge and I. Gruais, *Asymptotics of arbitrary order for a thin elastic clamped plate.* II: *Analysis of the boundary layer terms*, Asymptot. Anal., 16 (1998), pp. 99–124.

[10] M. Dauge and I. Gruais, *Edge layers in thin elastic plates*, Comput. Methods Appl. Mech. Eng., 157 (1998), pp. 335–347.

[11] M. Dauge, I. Gruais, and A. Rössle, *The influence of lateral boundary conditions on the asymptotics in thin elastic plates* I: *clamped and simply supported plates*, prépublication 97-28, IRMAR, 1997; also available online at http://www.maths.univ-rennes1.fr/˜dauge/.

[12] M. Dauge and I. Gruais, *The influence of lateral boundary conditions on the asymptotics in thin elastic plates* II: *frictional, sliding edge and free plates*, prépublication 97-29, IRMAR, 1997; also available online at http://www.maths.univ-rennes1.fr/˜dauge.

[13] P. Destuynder, *Sur une Justification des Modèles de Plaques et de Coques par les Méthodes Asymptotiques*, thèse d'Etat, Université Pierre et Marie Curie, Paris, 1980.

[14] P. Destuynder, *Une Théorie Asymptotique des Plaques Minces en Élasticité Linéaire.*, RMA 2, Masson, Paris, 1986.

[15] K. O. Friedrichs and R. F. Dressler, *A boundary-layer theory for elastic plates.*, Comm. Pure Appl. Math., 14 (1961), pp. 1–33.

[16] A. L. Gol'denveizer, *Derivation of an approximate theory of bending of a plate by the method of asymptotic integration of the equations of the theory of elasticity.*, Prikl. Matem. Mekhan., 26 (1962), pp. 668–686; J. Appl. Maths. Mech., 26 (1964) pp. 1000–1025 (in English).

[17] R. D. Gregory and F. Y. Wan, *Decaying states of plane strain in a semi-infinite strip and boundary conditions for plate theory*, J. Elasticity, 14 (1984), pp. 27–64.

[18] G. Kirchhoff, *Über das Gleichgewicht und die Bewegung einer elastischen Scheibe*, J. Reine Angew. Math., 40 (1850), pp. 51–58.

[19] V. G. Maz'ya, S. A. Nazarov, and B. A. Plamenevskii, *Asymptotische Theorie Elliptischer Randwertaufgaben in Singulär Gestörten Gebieten* II, Mathematische Monographien, Band 83, Akademie Verlag, Berlin, 1991.

[20] A. Mielke, *On the justification of plate theories in linear elasticity theory using exponential decay estimates*, J. Elasticity, 38 (1995), pp. 165–208.

[21] D. Morgenstern, *Herleitung der Plattentheorie aus der dreidimensionalen Elastizitästheorie.*, Arch. Rat. Mech. Anal., 4 (1959), pp. 145–152.

[22] S. A. Nazarov, *On the accuracy of asymptotic approximations for longitudinal deformation of a thin plate*, Math. Model. Numer. Anal., 30 (1996), pp. 185–213.

[23] S. A. Nazarov and B. A. Plamenevskii, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, in de Gruyter Exp. Math., de Gruyter, Berlin, 1994.

[24] S. A. Nazarov and I. S. Zorin, *Edge effect in the bending of a thin three-dimensional plate.*, Prikl. Matem. Mekhan., 53 (1989), pp. 642–650; J. Appl. Maths. Mechs., 53 (1989) 500–507 (in English).

[25] J. Nečas and I. Hlavaček, *Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction*, Elsevier Scientific Publishing Company, Amsterdam, 1981.

[26] P. F. Papkovich, *Stroitel'naia Mekhanika Korablia, Part* II *(Structural Mechanics of Ships)*, Sudpromgiz Publishers, Russia, 1941.

[27] J. C. Paumier, *Existence and convergence of the expansion in the asymptotic theory of elastic thin plates.*, RAIRO Modél. Math. Anal. Numér., 25 (1997), pp. 371–391.

[28] J. C. Paumier and A. Raoult, *Asymptotic consistency of the polynomial approximation in the linearized plate theory, application to the Reissner-Mindlin model*, in Elasticité, Viscoélasticité et Contrôle Optimal: Huitièmes Entretiens du Centre Jacques Cartier, SMAI, ESAIM Proceedings, 2, 1997, pp. 203–213.

[29] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of the function space.*, Quart. Appl. Math., 5 (1947), pp. 241–269.

[30] A. Rössle, M. Bischoff, W. L. Wendland, and E. Ramm, *On the mathematical foundation of the* $(1,1,2)$*-plate model*, Internat. J. Solids Structures, 36 (1999), pp. 2143–2168.

[31] B. A. Shoikhet, *An energy identity in physically nonlinear elasticity and error estimates of the plate equations*, Prikl. Matem. Mekhan., 40 (1976), pp. 317–326; J. Appl. Maths. Mechs., 40 (1976), pp. 291–301 (in English).

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 346–374

ON TRANSITIONS TO STATIONARY STATES IN A MAXWELL–LANDAU–LIFSCHITZ–GILBERT SYSTEM[*]

PATRICK JOLY[†], ALEXANDER KOMECH[‡], AND OLIVIER VACUS[§]

Abstract. In this paper we consider Maxwell's equations together with a dissipative nonlinear magnetic law, the Landau–Lifschitz–Gilbert equation, and we study long-time asymptotics of solutions in the 1D case in an infinite domain of propagation. We prove long-time convergence to zero of the electromagnetic field in a Fréchet topology defined by local energy seminorms: this corresponds to the local energy decay. We then introduce the set of stationary states for the Landau–Lifschitz–Gilbert equation and prove that it corresponds to the attractor set for the distribution of magnetization whose presence is one of the characteristics of ferromagnetic media.

Key words. Maxwell's equations, Landau–Lifchitz–Gilbert law, local energy decay, Liapunov theory, long-time asymptotics

AMS subject classifications. 15A15, 15A09, 15A23

PII. S0036141097329949

1. Introduction.

Ferromagnetic materials possess a spontaneous magnetization whose interaction with the magnetic field provides to this type of medium interesting absorbing properties with respect to electromagnetic waves. That is why the use of such materials as absorbing coatings for scatterers is of real importance for stealth technology. The present paper is a contribution to the mathematical theory of electromagnetic scattering by such objects. One of the main characteristics of ferromagnetic materials lies in the fact that their constitutive law, namely, the relationship between the magnetic field H and the magnetization M, is nonlinear and nonlocal with respect to time. This equation is the Landau–Lifschitz–Gilbert (LLG) equation that can be written pointwise in the form

$$\dot{M} \;=\; \gamma H_T \times M + \frac{\alpha}{|M|} M \times \dot{M}, \tag{1.1}$$

where H_T is the total magnetic field defined as

$$H_T \;=\; H + H_s + H_a(M), \tag{1.2}$$

with each of these contributions being defined as

$$\left\{ \begin{array}{lll} H & & \text{is the magnetic field;} \\ H_s \;=\; H_s(\mathbf{x}) & & \text{is an exterior static field (given);} \\ H_a(M) \;=\; -K\;P(M) & & \text{is a field of anisotropy.} \end{array} \right. \tag{1.3}$$

In (1.3), $\mathbf{x} = (x, y, z)$ denotes the space variable, t denoting time; K denotes a positive coefficient, constant in time but that may depend on $\mathbf{x}$; and $P(M)$ is the orthogonal

[*] Received by the editors November 12, 1997; accepted for publication January 23, 1998; published electronically December 21, 1999. This research was partially supported by INRIA (Project "Ondes"), by French-Russian A.M. Liapunov Center of Moscow State University, and by RFBR research grant 96-01-00527.

http://www.siam.org/journals/sima/31-2/32994.html

[†] INRIA Rocquencourt, 78 153 Le Chesnay Cédex, France (patrick.joly@inria.fr).

[‡] Department of Mechanics and Mathematics, Moscow State University, Moscow 119899, Russia (komech@facet.inria.msu.ru).

[§] Dassault Aviation, DPR-DESA, 78 quai Marcel Dassault, 92214 St. Cloud Cedex, France (olivier.vacus@dassault-aviation.fr).

projection in $\mathbb{R}^3$ on the plane orthogonal to some unit vector $\mathbf{p}$, called the easy axis (this direction, which is linked to the crystallic structure of the material, may also depend on $\mathbf{x}$):

$$P(M) = M - (p \cdot M)p\ . \tag{1.4}$$

Let us mention that in (1.1), γ, the gyromagnetic factor, is a universal constant while α, the damping factor, is a phenomenologic coefficient which depends on $\mathbf{x}$. Therefore, for our applications, a propagation medium will be determined by H_s, α, K, as function of $\mathbf{x}$ (and by the initial distribution of magnetization M_0; see (1.7)).

Here we are interested in the coupling of (1.1) with Maxwell's equations in some domain Ω (typically an exterior domain, if one thinks of applications to scattering problems) with boundary $\Gamma = \partial\Omega$. We assume that space and time variables are scaled in such a way that the constants γ, ε_0 (the electric permittivity) and μ_0 (the magnetic permeability) can be taken equal to 1. One then has to solve

$$\left\{\begin{array}{l} \dot{E}(\mathbf{x},t) - \text{curl } H(\mathbf{x},t) = 0, \\ \dot{H}(\mathbf{x},t) + \text{curl } E(\mathbf{x},t) + \dot{M}(\mathbf{x},t) = 0, \\ \dot{M}(\mathbf{x},t) = H_T(\mathbf{x},t) \times M(\mathbf{x},t) + \dfrac{\alpha}{|M(\mathbf{x},t)|} M(\mathbf{x},t) \times \dot{M}(\mathbf{x},t), \end{array}\right. \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{1.5}$$

with a perfectly conducting boundary condition on Γ with unit normal vector n

$$E \times n\ |_\Gamma =\ 0, \tag{1.6}$$

and initial conditions on $\mathbb{R}_-$:

$$E(\mathbf{x}, t=0)\ =\ E_0(\mathbf{x}),\ H(\mathbf{x}, t=0)\ =\ H_0(\mathbf{x}),\ M(\mathbf{x}, t=0)\ =\ M_0(\mathbf{x}). \tag{1.7}$$

From a mathematical point of view, even the existence and uniqueness result for system (1.5) appears to be a very difficult question (see [4], [5]). Another natural question that we wish to address here is the following: Is it possible to describe the asymptotic behavior of the solution of system (1.9) for large time? In the case of linear materials, the answer to this question has been known for a long time (see [9], [10], [11]): the electric and magnetic fields tend locally to 0. This result is known as the local energy decay. A more subtle question is an estimate of the rate of decay; this question is closely related to the geometry of the obstacle [18].

In the case of nonlinear media, there are much fewer results in that direction (see [16], [17]). Our goal in this paper is to establish a result analogous to the local energy decay in a simplified 1*D model problem.* More precisely we assume that all the unknowns are functions of only one space variable x (i.e., we consider the propagation of plane waves). The curl operator is then defined by

$$\text{curl } H(x,t) = \left(0, -\frac{\partial H_z}{\partial x}, \frac{\partial H_y}{\partial x}\right),\ \text{curl } E(x,t) = \left(0, -\frac{\partial E_z}{\partial x}, \frac{\partial E_y}{\partial x}\right). \tag{1.8}$$

We assume that the propagation medium is the half-space $x < 0$ and apply the perfectly conducting boundary condition at $x = 0$ ($e_x = (1,0,0)$):

$$E \times e_x\ =\ 0.$$

We also assume that the support of the initial magnetization M_0, which defines the ferromagnetic layer (see 2.6), is compact:

$$\text{supp } M_0\ \subset\ ]-a, 0].$$

Then, by a principle of reflection (or image principle), the analysis of (1.5), (1.6), and (1.7) can be reduced to the analysis of the pure Cauchy problem on the whole line:

$$\left\{\begin{array}{l} \dot{E}(x,t) - \text{curl } H(x,t) = 0, \\ \dot{H}(x,t) + \text{curl } E(x,t) + \dot{M}(x,t) = 0, \\ \dot{M}(x,t) = H_T(x,t) \times M(x,t) + \dfrac{\alpha}{|M(x,t)|} M(x,t) \times \dot{M}(x,t), \end{array}\right. \qquad x \in \mathbb{R},\ t > 0, \tag{1.9}$$

provided that the new initial data M_0, E_0, H_0 are appropriate extensions of the original ones:

$$\left\{\begin{array}{lcl} M_0(-x) & = & M_0(x), \\ E_0(-x) & = & \Pi_\perp\, E_0(x), \\ H_0(-x) & = & \Pi_{\|}\, H_0(x), \end{array}\right. \tag{1.10}$$

where the operators $\Pi_\perp$ and $\Pi_{\|}$ are defined for any field $A(A_x, A_y, A_z)$ by

$$\left|\begin{array}{lcl} \Pi_\perp(A_x, A_y, A_z) & = & (A_x, -A_y, -A_z), \\ \Pi_{\|}(A_x, A_y, A_z) & = & (-A_x, A_y, A_z). \end{array}\right. \tag{1.11}$$

Remark 1.1. Concerning the longitudinal components, equations (1.9) imply

$$\begin{array}{c} E_x(x,t) = E_x^0(x), \\ H_x(x,t) + M_x(x,t) = H_x^0(x) + M_x^0(x). \end{array} \tag{1.12}$$

We see here that the component E_x is constant in time, while convergence results on M_x yield results on H_x.

The outline of this article is as follows. In section 2, after having recalled some known results about weak and strong solutions of the 1D scattering problem, we state the main results of the paper, namely, Theorems A and A$'$ for the local energy decay (respectively, for weak and strong solutions) and Theorem B on long-time asymptotics of the magnetization M. Sections 3 and 4 are, respectively, devoted to the proof of Theorems A and A$'$. In section 5, we give some intermediate results about the LLG equation seen as an ordinary differential equation; these results are preparatory for section 6, in which we prove Theorem B. Finally, section 7 is devoted to some additional remarks and comments.

2. Statement of the main results.

2.1. Overview of known results in the 1D case. As we said above, existence and uniqueness results for system (1.9) are very difficult in dimension 3. In the case of the 1D model, the problem is easier and can be handled via a fixed point theorem. Let us summarize the main existence and uniqueness results from [7].

We introduce the phase space V for the system (1.9). Let $L^p = [L^p(\mathbb{R})]^3$ for $1 \le p \le \infty$, $L^{2,\infty}$ be the Banach space $L^2 \cap L^\infty$ with the norm

$$\|M\|_{L^{2,\infty}} = \|M\|_{L^2} + \|M\|_{L^\infty}, \tag{2.1}$$

and let $H(\text{curl})$ denote the Hilbert space $\{u \in L^2 \,/\, \text{curl } u \in L^2\}$ with the norm

$$\|u\|_{\text{curl}}^2 = \|\text{curl } u\|_{L^2}^2 + \|u\|_{L^2}^2. \tag{2.2}$$

Definition 2.1. *Let V be the Banach space*

$$V \;=\; \Big\{(E, H, M) \in H(\text{curl}) \times H(\text{curl}) \times L^{2,\infty};\ H_x \in L^\infty(\mathbb{R})\Big\}$$

equipped with the norm

$$\|(E,H,M)\|_V = \|E\|_{\text{curl}} + \|H\|_{\text{curl}} + \|M\|_{L^{2,\infty}} + \|H_x\|_{L^\infty(\mathbb{R})}. \tag{2.3}$$

Remark 2.2. The x component H_x of the magnetic field plays a particular role because of the particular form of the curl operator in the 1D case.

Definition 2.3. *A function* $Y(t) = (E(x,t), H(x,t), M(x,t)) \in C^0(0,\infty;V)$ *is a global strong solution to the system* (1.9) *if*

$$(E,H) \in C^1(0,\infty;L^2) \cap C^0(0,\infty;H^1) \text{ and } M \in C^1(0,\infty;L^{2,\infty}) \cap C^2(0,\infty;L^2) \tag{2.4}$$

and all equations in (1.9) *hold in the sense of distributions.*

One then shows the following theorem.

Theorem 2.4. *Let the following assumptions hold:*

$$\begin{cases} \bullet\ \alpha(x), K(x) \in L^\infty(\mathbb{R}),\ p(x) \in L^\infty; \\ \bullet\ H_s(x) \in L^{2,\infty}, \\ \bullet\ (E^0(x), H^0(x), M^0(x)) \in V. \end{cases} \tag{2.5}$$

Then the Cauchy problem (1.9) *admits a unique global strong solution* (E,H,M)*, which, moreover, satisfies*

$$|M(x,t)| = |M_0(x)| \quad a.e.\ x \in \mathbb{R}\ \forall t \geq 0, \tag{2.6}$$

$$\frac{d}{dt}\mathcal{E}(E,H,M) + \int_{\mathbb{R}} \frac{\alpha}{|M|} \left|\dot{M}\right|^2 dx = 0, \tag{2.7}$$

where $\mathcal{E}(E,H,M)$ *denotes the electromagnetic energy defined by*

$$\mathcal{E}(E,H,M) = \frac{1}{2}\int_{\mathbb{R}} \left[|E|^2 + |H|^2 + K|P(M)|^2 + |H_s - M|^2\right] dx\,. \tag{2.8}$$

Remark 2.5. For any strong solution, the electromagnetic energy is a function of class C^1 with respect to time.

One has to make precise the sense of the integral of $\frac{\alpha}{|M|}|\dot{M}|^2$. In fact, from the LLG equation, we have

$$\dot{M}(t) - \alpha\frac{M}{|M|} \times \dot{M}(t) = \gamma H_T(M(t)) \times M(t);$$

hence we deduce, via Pythagoras's theorem, that

$$(1+\alpha^2)\,|\dot{M}(t)|^2 = \gamma^2\,|H_T(M) \times M|^2\,. \tag{2.9}$$

This observation leads to the following definition.

Definition 2.6. *For the strong solution* (E,H,M) *to system* (1.9)*, we set*

$$\frac{\alpha}{|M|}\left|\dot{M}\right|^2 = \gamma^2 \frac{\alpha}{1+\alpha^2}\frac{|H_T \times M|^2}{|M|}, \tag{2.10}$$

which is finite since the function $M \mapsto \frac{|H_T \times M|^2}{|M|}$ *can be continuously extended by* 0 *for* $M = 0$. *We have the estimate*

$$\int_{\mathbb{R}} \frac{\alpha}{|M|}\left|\dot{M}\right|^2 dx \leq \gamma^2 \int_{\mathbb{R}} \frac{\alpha}{1+\alpha^2}|M||H_T|^2\,dx \tag{2.11}$$

which makes sense since one easily checks that $M \in (L^\infty)^3$ *and* $H_T \in (L^2)^3$ *for any time.*

Proof. See [7] for a complete proof. We just explain below how to obtain the two estimates of Theorem 2.4. Concerning (2.6), the product of the LLG equation with M shows that $M \cdot \dot{M} = 0$; we deduce that

$$|M(x,t)| \;=\; |M_0(x)|, \quad \text{a.e. } x \in \mathbb{R} \;\, \forall t \geq 0.$$

For (2.7), from Maxwell's equations we get

$$\left\{ \begin{array}{lcl} \dot{E} \cdot E - \mathrm{curl} H \cdot E &=& 0, \\ \dot{H} \cdot H + \mathrm{curl} E \cdot H &=& -\dot{M} \cdot H. \end{array} \right. \tag{2.12}$$

Summing these two equalities and integrating over $\mathbb{R}$ leads, after integration by parts, to the following identity:

$$\frac{d}{dt} \left\{ \frac{1}{2} \int_{\mathbb{R}} \left(|E|^2 + |H|^2 \right) \, dx \right\} \;=\; - \int_{\mathbb{R}} \dot{M} \cdot H \, dx. \tag{2.13}$$

We now use the LLG equation:

$$\dot{M} \;=\; \gamma H_T \times M + \frac{\alpha}{|M|} M \times \dot{M}. \tag{2.14}$$

Taking the scalar product with H_T, and using the notation $(\cdot,\cdot,\cdot)$ for the mixed product in $\mathbb{R}^3$, we get

$$\dot{M} \cdot H_T = \frac{\alpha}{|M|} \left(M, \dot{M}, H_T \right).$$

Taking the scalar product of the LLG equation by $\dot{M}$ gives

$$\left| \dot{M} \right|^2 = \gamma \left(\dot{M}, H_T, M \right).$$

Therefore, eliminating the mixed product gives

$$\dot{M} \cdot H_T = \frac{\alpha}{\gamma |M|} \left| \dot{M} \right|^2. \tag{2.15}$$

(The meaning of the right-hand side of this expression has to be understood in the sense we made precise in Definition 2.6.) Now, by definition of H_T, we have

$$\dot{M} \cdot H_T = \dot{M} \cdot H - K P\left(M\right) \cdot P\left(\dot{M}\right) + H_s \cdot \dot{M}.$$

Using the fact that $|M(x,t)| = |M_0(x)|$, we note that

$$\left\{ \begin{array}{l} H_s \cdot \dot{M} = -\dfrac{1}{2} \dfrac{\partial}{\partial t} |H_s - M|^2, \\ -K P\left(M\right) \cdot P\left(\dot{M}\right) = -\dfrac{1}{2} \dfrac{\partial}{\partial t} \left[K |P(M)|^2 \right]. \end{array} \right. \tag{2.16}$$

Therefore, we have

$$\dot{M} \cdot H_T = \dot{M} \cdot H - \frac{1}{2} \frac{\partial}{\partial t} \left[|H_s - M|^2 + K |P(M)|^2 \right]; \tag{2.17}$$

that is to say, using (2.15),

$$-\dot{M}\cdot H = -\frac{\alpha}{\gamma|M|}\left|\dot{M}\right|^2 - \frac{1}{2}\frac{\partial}{\partial t}\left[|H_s - M|^2 + K|P(M)|^2\right]. \tag{2.18}$$

Plugging (2.18) into (2.13) leads to the energy identity

$$\frac{d}{dt}\left\{\frac{1}{2}\int_{\mathbb{R}}\left(|E|^2 + |H|^2 + |H_s - M|^2 + K|P(M)|^2\right)dx\right\} \tag{2.19}$$
$$+\frac{1}{\gamma}\int_{\mathbb{R}}\frac{\alpha}{|M|}\left|\dot{M}\right|^2 dx = 0. \qquad \square$$

Thanks to a priori estimates (2.6) and (2.7), one is also able to obtain a theorem for weak solutions (i.e., less regular solutions) under weaker assumptions on the initial data; see the following definition.

DEFINITION 2.7. *A function* $Y(t) = (E(x,t), H(x,t), M(x,t)) \in C^0(0,\infty;V)$ *is a global weak solution to the system* (1.9) *if and only if*

(i) $M \in C^1(0,\infty;L^{2,\infty})$;

(ii) *for any* $(\varphi,\psi) \in V \times V$ *where* V *is the space of test fields*

$$V = \left\{\varphi \in C^1(H(\mathrm{curl}))\ /\ \mathrm{supp}\ (\varphi)\ \textit{is compact}\right\},$$

$$\left\{\begin{array}{rcl} \displaystyle\iint(\dot{\varphi}\cdot E + \mathrm{curl}\ \varphi\cdot H)\ dxdt & = & \displaystyle -\int E_0\cdot\varphi(\cdot,0)dx, \\ \displaystyle\iint(\dot{\psi}\cdot H - \mathrm{curl}\ \psi\cdot E)\ dxdt & = & \displaystyle -\int H_0\cdot\psi(\cdot,0)dx\ -\ \iint\dot{M}\cdot\psi\ dxdt; \end{array}\right.$$

(iii) *for almost every* $(x,t) \in \mathbb{R}\times\mathbb{R}^+$,

$$\left\{\begin{array}{l} \dot{M}(x,t) = H_T(x,t)\times M(x,t) + \dfrac{\alpha}{|M(x,t)|}M(x,t)\times\dot{M}(x,t), \\ M(x,0) = M_0(x). \end{array}\right.$$

THEOREM 2.8. *Assume that*

$$(E_0, H_0, M_0)\ \in\ (L^2)^3\times(L^2)^3\times(L^{2,\infty})^3;$$

then system (1.9) *admits a unique global weak solution which satisfies*

$$\int_0^\infty\int_{\mathbb{R}}\frac{\alpha}{|M|}|\dot{M}|^2\,dx\,dt\ <\ +\infty. \tag{2.20}$$

Moreover, a.e. $x \in \mathbb{R}$, $t \mapsto M(x,t)$ *belongs to* $C^0(\mathbb{R})$.

Remark 2.9. From (2.20) and Fubini's theorem, we deduce that

$$\text{a.e. } x \in \mathbb{R}, \quad \frac{|\dot{M}(x,t)|^2}{|M_0(x)|} \in L^1(\mathbb{R}).$$

Therefore, function $t \mapsto M(x,t)$ is in $H^1(\mathbb{R}_+) \subset C^0(\mathbb{R}_+)$.

2.2. Main results of the paper. Our first main theorem concerns the convergence to 0 in Fréchet topology of the transverse components of the electromagnetic fields. In the following, we denote the transverse components $E_{\|}(x,t) = (E_y(x,t), E_z(x,t))$, $H_{\|}(x,t) = (H_y(x,t), H_z(x,t))$, and so on.

THEOREM A. *Let all assumptions* (2.5) *hold. Assume, moreover, that*

$$M_0(x) \ = \ 0 \quad \textit{for } |x| > a \tag{2.21}$$

and

$$\exists\, \alpha_* > 0 \textit{ such that } \ \alpha(x) \geq \alpha^*, \ \textit{a.e.} \quad x \in [-a;a]. \tag{2.22}$$

Then, for the global solution to the Cauchy problem (1.9):

(i) *For almost every* $x \in \mathbb{R}$

$$\int_0^\infty (|E_{\|}(x,t)|^2 + |H_{\|}(x,t)|^2)dt < \infty. \tag{2.23}$$

(ii) *For every* $R > 0$

$$\int_{|x|<R} (|E_{\|}(x,t)|^2 + |H_{\|}(x,t)|^2)dx \to 0 \textit{ as } t \to \infty. \tag{2.24}$$

Our second result is a variant of Theorem A. We prove that, provided additional regularity assumptions on the initial data, we have convergence to zero of the transverse electromagnetic field, not only in the local energy norm, but also uniformly in any compact set.

THEOREM A′. *Let the assumptions of Theorem* A *hold (in particular* $(E_0, H_0) \in (H(\text{curl}))^2$. *Then*

$$\lim_{t\to+\infty} \int_{-a}^{a} |\dot{M}(x,t)|^2 dx = 0, \tag{2.25}$$

and, for every $R > 0$,

$$\lim_{t\to+\infty} \int_{-R}^{R} \left(\left| \frac{\partial E_{\|}}{\partial x}(x,t) \right|^2 + \left| \frac{\partial H_{\|}}{\partial x}(x,t) \right|^2 \right) dx = 0, \tag{2.26}$$

$$\lim_{t\to+\infty} \sup_{|x|\leq R} \Big\{ |E_{\|}(x,t)| + |H_{\|}(x,t)| \Big\} = 0. \tag{2.27}$$

Our other results concern the asymptotic behavior of the magnetization $M(x,t)$ (and thus of the longitudinal magnetic field). We first need to introduce a nondegeneracy assumption. Let us define

$$\tilde{H}_T(x,M) \ = \ H_s + K(p \cdot M)p - (e_x \cdot M)e_x, \tag{2.28}$$

where the dependence of $\tilde{H}_T(x,M)$ with respect to x appears in the dependence of H_s, K, and p. We introduce the assumption, for any x in $\mathbb{R}$,

$$(\mathcal{H}_x) \ : \ (\forall M \in \mathbb{R}^3, \ \exists\, \lambda(x,M) \in \mathbb{R}, \ \tilde{H}_T(M) = \lambda(x,M)e_x\,) \ \Rightarrow \ (\forall M \in \mathbb{R}^3, \ \lambda(x,M) \neq 0\,).$$

Remark 2.10. Assumption $(\mathcal{H}_x)$ is not satisfied if and only if the following two properties are satisfied:

(i) H_s and p are collinear to e_x ;
(ii) there exists $M \in \mathbb{R}^3$ such that $(H_s + (K-1)M) \cdot e_x \ = \ 0$.
For any x, we introduce the set

$$\mathcal{Z}(x, M_0) \ = \ \Big\{ M \in \mathbb{R}^3 \text{ such that } \tilde{H}_T(x, M) \times M = 0 \text{ and } |M| = |M_0(x)| \Big\}.$$

This set will be identified in section 5 as the set of stationary states of some unperturbed LLG equation at point x that are possible limits for $M(x,t)$ with the initial data $M_0(x)$. It will be proved to be a finite set containing between two and six elements (see Theorem 5.2).

Our first result for M concerns weak solutions.

Theorem B. *Let the assumptions of Theorem* A *be satisfied. Assume that* $(\mathcal{H}_x)$ *holds almost everywhere in* x. *Then*

$$M(x,t) \to \mathcal{Z}(x, M_0) \ \textit{as}\ t \to \infty \ \textit{for a.e.}\ x \in \mathbb{R}. \tag{2.29}$$

Remark 2.11. With the set $\mathcal{Z}(x, M_0)$ being discrete, (2.29) means that for a.e. x, there exists $M_\infty \in \mathcal{Z}(x, M_0)$ such that

$$M(x,t) \to M_\infty \text{ as } t \to \infty.$$

In other words,

$$\mathcal{M} \ = \ \{\, M \in L^\infty([-a,a]) \,/\, \text{for a.e. } x \in [-a,a],\ M(x) \ = \ M_\infty \in \mathcal{Z}(x, M_0) \,\}$$

is an infinite-dimensional attractor for $M(x,t)$.

Our last result is a variant of Theorem B for strong solutions.

Corollary 2.12. *Let the assumptions of Theorem* A′ *be satisfied. Assume that* $(\mathcal{H}_x)$ *holds everywhere. For every* x, *there exists* $M_\infty(x) \in \mathcal{Z}(x, M_0)$ *such that*

$$\lim_{t\to\infty} |M(x,t) - M_\infty(x)| \ = \ 0. \tag{2.30}$$

Remark 2.13. We show with Theorem A′ that the transverse magnetic field $H_\parallel$ converges in time to 0 uniformly in space. It must be emphasized that this property is not strong enough to ensure that the convergence of the magnetization M is also uniform in space. Section 7 will be devoted to a counterexample and some comments about this assertion.

3. Proof of Theorem A: L^2 bounds of the transverse electromagnetic field for weak solutions. The first two equations of (1.9) read

$$\begin{cases} \dot{E}_x = 0, & \dot{E}_y + \dfrac{\partial H_z}{\partial x} = 0, & \dot{E}_z - \dfrac{\partial H_y}{\partial x} = 0, \\[2ex] \dot{H}_x + \dot{M}_x = 0, & \dot{H}_y - \dfrac{\partial E_z}{\partial x} + \dot{M}_y = 0, & \dot{H}_z + \dfrac{\partial E_y}{\partial x} + \dot{M}_z = 0. \end{cases} \tag{3.1}$$

We easily deduce that

$$\begin{cases} L_+(E_y + H_z) & = - & \dot{M}_z, \\ L_-(E_y - H_z) & = & \dot{M}_z, \\ L_+(E_z - H_y) & = & \dot{M}_y, \\ L_-(E_z + H_y) & = - & \dot{M}_y, \end{cases} \tag{3.2}$$

where we have introduced the two transport operators $L_\pm = \frac{\partial}{\partial t} \pm \frac{\partial}{\partial x}$. Considering $\dot{M}$ as known, we can solve explicitly (3.1) using the method of characteristics. After some algebraic manipulations, we end up with the following formulas:

$$(3.3)\quad \left\{ \begin{aligned} E_\|(x,t) &= \tfrac{1}{2}\{E_\|^0(x+t) + E_\|^0(x-t)\} + \tfrac{J}{2}\{H_\|^0(x+t) - H_\|^0(x-t)\} \\ &\quad - \tfrac{J}{2}\left\{\int_{\Gamma_+(x,t)} \dot{M}_\|(y,s)d\sigma - \int_{\Gamma_-(x,t)} \dot{M}_\|(y,s)d\sigma\right\}, \\ H_\|(x,t) &= \tfrac{1}{2}\{H_\|^0(x+t) + H_\|^0(x-t)\} - \tfrac{J}{2}\{E_\|^0(x+t) - E_\|^0(x-t)\} \\ &\quad - \left\{\int_{\Gamma_+(x,t)} \dot{M}_\|(y,s)d\sigma + \int_{\Gamma_-(x,t)} \dot{M}_\|(y,s)d\sigma\right\}, \end{aligned} \right.$$

where

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

$\Gamma_\pm(x,t)$ is the curve $\Gamma_\pm(x,t) = (x,t) - D_\pm$, with $D_\pm = \{(x,t)/x \pm t = 0\}$ (see Figure 3.1), and σ is the curvilinear abscissa along $\Gamma_\pm(x,t)$. Therefore, introducing

$$F(x,t) = \begin{pmatrix} E_\|(x,t) \\ H_\|(x,t) \end{pmatrix},$$

we get (note that $d\sigma = \sqrt{2}dy$, $\|J\| = 1$, and $|\dot{M}_\|| \le |\dot{M}|$)

$$(3.4)\qquad |F(x,t)| \le \frac{\sqrt{2}}{2}\int_{\Gamma^a(x,t)} |\dot{M}(y,s)|dy + \frac{1}{2}\{|F^0(x+t)| + |F^0(x-t)|\}$$

with $\Gamma^a(x,t) = \{(y,s) \in \Gamma_+(x,t) \cup \Gamma_-(x,t) :\ |y| < a\}$ (see Figure 3.1).

Then, the Cauchy–Schwarz inequality yields the estimate ($\int_{\Gamma^a(x,t)} dy = 2a$)

$$(3.5)\quad |F(x,t)| \le \sqrt{a}\left(\int_{\Gamma^a(x,t)} |\dot{M}(y,s)|^2 dy\right)^{\frac{1}{2}} + \frac{1}{2}\{|F^0(x+t)| + |F^0(x-t)|\}.$$

Integrating this inequality between 0 and T, and noticing that

$$K(x,T) \equiv \bigcup_{t=0}^{T} \Gamma^a(x,t)$$

is such that $K(x,T) \subset [-a,a] \times [0,T]$, we get the estimate

$$\int_0^T |F(x,t)|^2 dt \ \le\ 2a\int_0^T \left(\int_{|y|<a} |\dot{M}(y,s)|^2 dy\right) ds \ +\ \frac{1}{2}\sum_\pm \int_0^T |F^0(x \pm t)|^2 dt,$$

which yields, after integrating over $[-a,a]$,

$$\left|\ \begin{aligned} \int_0^T \int_{-a}^{a} |F(x,t)|^2 dxdt \quad &\le \quad 4a^2 \int_0^T \left(\int_{|y|<a} |\dot{M}(y,s)|^2 dy\right) ds \\ &\quad + \quad \frac{1}{2}\sum_\pm \int_0^T \int_{-a}^{a} |F^0(x \pm t)|^2 dt. \end{aligned}\right.$$

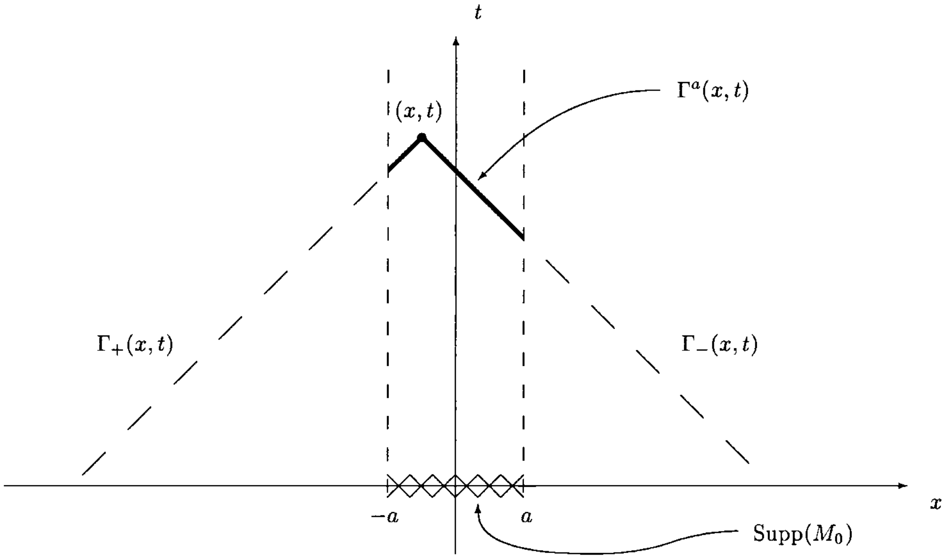


FIG. 3.1. *Characteristic lines and the 1D layer.*

Writing $|\dot M|^2 = \frac{|M|}{\alpha}\cdot\frac{\alpha|\dot M|^2}{|M|}$ and using the conservation of the norm of M and the assumption (2.22) about the damping function $\alpha(x)$, we get

$$\left|\begin{aligned}\int_0^T\int_{-a}^a |F(x,t)|^2 dxdt \;&\le\; \frac{4a^2}{\alpha_*}\|M_0\|_\infty\int_0^T\left(\int_{|y|<a}\frac{\alpha}{|M|}|\dot M(y,s)|^2dy\right)ds\\ &+\;\frac12\sum_{\pm}\int_0^T\int_{-a}^a|F^0(x\pm t)|^2dt.\end{aligned}\right. \tag{3.6}$$

Therefore, (2.5) and (2.20) imply (2.23). Besides, (3.4) can be rewritten as

$$\left|\begin{aligned}|F(x,t)| \;\le\; \frac{\sqrt2}{2}\int_{\Gamma_a^+(x,t)}|\dot M(y,s)|\,ds \;&+\; \frac{\sqrt2}{2}\int_{\Gamma_a^-(x,t)}|\dot M(y,s)|\,ds\\ &+\;\frac12\left\{|F^0(x+t)|\,+\,|F^0(x-t)|\right\},\end{aligned}\right.$$

where $\Gamma_a^+ = \Gamma_+\cap\Gamma^a$ and $\Gamma_a^- = \Gamma_-\cap\Gamma^a$. This yields, via the Cauchy–Schwarz inequality,

$$\left|\begin{aligned}|F(x,t)| \;\le\; \frac{\sqrt{2a}}{2}\left(\int_{\Gamma_a^+(x,t)}|\dot M(y,s)|^2ds\right)^{\frac12} \;&+\; \frac{\sqrt{2a}}{2}\left(\int_{\Gamma_a^-(x,t)}|\dot M(y,s)|^2ds\right)^{\frac12}\\ &+\;\frac12\left(|F^0(x+t)|+|F^0(x-t)|\right).\end{aligned}\right.$$

Therefore

$$\left|\begin{aligned}|F(x,t)|^2 \;\le\; 2a\left(\int_{\Gamma_a^+(x,t)}|\dot M(y,s)|^2ds \;+\; \int_{\Gamma_a^-(x,t)}|\dot M(y,s)|^2ds\right)&\\ +\;|F^0(x+t)|^2+|F^0(x-t)|^2.&\end{aligned}\right. \tag{3.7}$$

After having remarked that, for $t > R + a$,

$$\bigcup_{x=-R}^{R} \Gamma_a^{\pm}(x,t) \subset [-a,a] \times [t-(R+a), t],$$

we obtain

$$(3.8) \quad \left| \begin{array}{rcl} \displaystyle\int_{-R}^{R} |F(x,t)|^2 dx & \leq & \displaystyle 4a \int_{t-(R+a)}^{t} \int_{-a}^{a} |\dot{M}(y,s)|^2 dyds \\ & & \displaystyle + \int_{-R}^{R} \left(|F^0(x+t)|^2 + |F^0(x-t)|^2 \right) dx \end{array} \right.$$

which leads to

$$(3.9) \quad \left| \begin{array}{rcl} \displaystyle\int_{-R}^{R} |F(x,t)|^2 dx & \leq & \displaystyle \frac{4a\|M_0\|_\infty}{\alpha_*} \int_{t-(R+a)}^{t} \int_{-a}^{a} \frac{\alpha}{|M|} |\dot{M}(y,s)|^2 dyds \\ & & \displaystyle + \int_{-R-t}^{R-t} |F^0(x)|^2 dx + \int_{-R+t}^{R+t} |F^0(x)|^2 dx. \end{array} \right.$$

By (2.5) and (2.7),

$$\int_{-\infty}^{+\infty} |F^0(x)|^2 dx < +\infty \quad \text{and} \quad \int_0^{\infty} \int_{-a}^{a} \frac{\alpha}{|M|} |\dot{M}|^2 dyds < +\infty.$$

Therefore, (3.9) implies (2.24). This concludes the proof. □

4. Proof of Theorem A′: Uniform bounds of the transverse electromagnetic field for strong solutions.

The proof of Theorem A′ relies on a technical lemma.

LEMMA 4.1. *The solution of the system satisfies*

$$(4.1) \qquad \int_0^{\infty} \int_{-a}^{a} \frac{\alpha}{|M_0(x)|} |\ddot{M}(x,t)|^2 dxdt < \infty.$$

Proof. We give a proof which supposes that the electromagnetic field (E, H) is slightly more regular in time than local strong solutions. However, all the forthcoming assumptions can be justified using the method of differential quotients (see [7] for details).

Nevertheless, our proof remains rather long and will be divided into four steps.

Step 1: *Estimates using the characteristics.* In what follows, we keep the notation of the proof of Theorem A. In particular, we set

$$(4.2) \qquad F(x,t) = \begin{bmatrix} E_{\|}(x,t) \\ H_{\|}(x,t) \end{bmatrix} \in \mathbb{R}^4.$$

First, we note that if we differentiate the system with respect to time, thanks to the linearity of Maxwell's equations, the relationship between $\ddot{M}$ and $\dot{E}$, $\dot{H}$ is exactly the same as the one between $\dot{M}$ and E, H. Therefore, reproducing the computations in the proof of Theorem A leads to the estimate

$$(4.3) \quad |\dot{F}(x,t)| \leq \sqrt{a} \left(\int_{\Gamma^a(x,t)} |\ddot{M}(y,s)| dy \right)^{\frac{1}{2}} + \frac{1}{2} \Big\{ |\dot{F}^0(x+t)| + |\dot{F}^0(x-t)| \Big\}.$$

Then, proceeding as in the proof of Theorem A, we easily get

$$\left|\begin{aligned} \int_0^T \int_{-a}^a |\dot F(x,t)|^2 dxdt \quad &\le \quad \frac{4a^2}{\alpha_*}\|M_0\|_\infty \int_0^T \left(\int_{|y|<a} \frac{\alpha}{|M|}|\ddot M(y,s)|^2 dy\right) ds \\ &+ \quad \frac{1}{2}\sum_{\pm}\int_0^T \int_{-a}^a |\dot F^0(x\pm t)|^2 dt. \end{aligned}\right. \tag{4.4}$$

Step 2: *Energy-like estimates.* We start from the linear Maxwell's equations after time derivation:

$$\left\{\begin{aligned} &\ddot E - \operatorname{curl} \dot H = 0, \\ &\ddot H + \operatorname{curl} \dot E = -\ddot M. \end{aligned}\right. \tag{4.5}$$

We multiply the first equation of (4.5) by $\dot E$, the second by $\dot H$, and add the two resulting equations. After integration over x, we easily get, using Green's formula and the fact that M is supported in $[-a,a]\times\mathbb{R}^+$,

$$\frac{d}{dt}\left(\frac{1}{2}\int (|\dot E|^2 + |\dot H|^2)dx\right) + \int_{-a}^a \dot H \ddot M dx = 0. \tag{4.6}$$

Now, we differentiate in time the LLG equation which leads to

$$\ddot M = \dot H \times M + K(p\cdot\dot M)p\times M + H_T \times \dot M + \frac{\alpha}{|M|} M \times \ddot M, \tag{4.7}$$

using the fact that $|M|$ is constant in time. Multiplying (4.7) successively by $\dot H$ and $\ddot M$ gives

$$\left|\begin{aligned} \ddot M \cdot \dot H \quad &= \quad K(p\cdot\dot M)(p, M\dot M) + (H_T, \dot M, \dot H) + \frac{\alpha}{|M|}(M, \ddot M, \dot H), \\ |\ddot M|^2 \quad &= \quad (\dot H, M, \ddot M) + K(p\cdot \dot M)(p, M, \ddot M) + (H_T, \dot M, \ddot M). \end{aligned}\right. \tag{4.8}$$

By an adequate linear combination of these two equalities, we eliminate the mixed product $(\dot H, M, \ddot M)$. This leads to

$$\ddot M \cdot \dot H = \frac{\alpha}{|M|}|\ddot M|^2 + (H_T, \dot M, \dot H) - \frac{\alpha}{|M|}(H_T, \dot M, \ddot M) \tag{4.9}$$

that we plug into (4.6) to obtain

$$\frac{d}{dt}\left[\frac{1}{2}\int (|\dot H|^2 + |\dot E|^2)dx\right] + \int_{-a}^a \frac{\alpha}{|M|}|\ddot M|^2 dx \ = \ \int_{-a}^a \left\{\frac{\alpha}{|M|}(H_T, \dot M, \ddot M) - (H_T, \dot M, \dot H)\right\} dx. \tag{4.10}$$

Let us use the bounds

$$\left|\begin{aligned} &|(H_T, \dot M, \ddot M)| \le \frac{1}{2}|\ddot M|^2 + \frac{1}{2}|H_T|^2|\dot M|^2, \\ &|(H_T, \dot M, \dot H)| \le \varepsilon |\dot H|^2 + \frac{1}{4\varepsilon}|H_T|^2|\dot M|^2, \end{aligned}\right.$$

where ε is an arbitrary strictly positive real number. We then obtain

$$\left| \begin{aligned} &\frac{d}{dt}\left(\frac{1}{2}\int(|\dot{H}|^2+|\dot{E}|^2)dx\right)+\frac{1}{2}\int_{-a}^{a}\frac{\alpha}{|M|}|\ddot{M}|^2dx \\ &\qquad\qquad \le \varepsilon\int_{-a}^{a}|\dot{H}|^2dx+\int_{-a}^{a}\left(\frac{\alpha}{2|M|}+\frac{1}{4\varepsilon}\right)|H_T|^2|\dot{M}|^2dx, \end{aligned}\right.$$

which we can integrate between 0 and T, to obtain

$$\tag{4.11} \left| \begin{aligned} &\frac{1}{2}\int(|\dot{H}(x,T)|^2+|\dot{E}(x,T)|^2)dx+\frac{1}{2}\int_{-a}^{a}\int_0^T\frac{\alpha}{|M|}|\ddot{M}|^2dxdt \\ &\quad \le \frac{1}{2}\int(|\dot{H}(x,0)|^2+|\dot{E}(x,0)|^2)dx+\varepsilon\int_{-a}^{a}\int_0^T|\dot{H}|^2dxdt \\ &\qquad +\int_{-a}^{a}\int_0^T\left(\frac{\alpha}{2|M|}+\frac{1}{4\varepsilon}\right)|H_T|^2|\dot{M}|^2dxdt. \end{aligned}\right.$$

Step 3: *Combination of the two estimates.* We use (1.12) in Remark 1.1 to write

$$\int_{-a}^{a}\int_0^T|\dot{H}|^2dxdt \ \le\ \int_{-a}^{a}\int_0^T|\dot{F}|^2dxdt\ +\ \int_{-a}^{a}\int_0^T|\dot{M}_x|^2dxdt.$$

We thus obtain, using (4.4) in (4.11),

$$\left| \begin{aligned} &\frac{1}{2}\int(|\dot{H}(x,T)|^2+|\dot{E}(x,T)|^2)dx+\frac{1}{2}\int_{-a}^{a}\int_0^T\left(1-\frac{8\varepsilon a^2}{\alpha_*}\|M_0\|_\infty\right)\frac{\alpha}{|M|}|\ddot{M}|^2dxdt \\ &\quad \le \frac{1}{2}\int(|\dot{H}(x,0)|^2+|\dot{E}(x,0)|^2)dx+\frac{\varepsilon}{2}\sum_{\pm}\int_{-a}^{a}\int_0^T|\dot{F}^0(x\pm t)|^2dxdt \\ &\qquad +\int_{-a}^{a}\int_0^T\left(\frac{\alpha}{2|M|}+\frac{1}{4\varepsilon}\right)|H_T|^2|\dot{M}|^2dxdt+\varepsilon\int_{-a}^{a}\int_0^T|\dot{M}_x|^2dxdt. \end{aligned}\right.$$

We observe that

$$\int_{-a}^{a}\int_0^T|\dot{F}^0(x\pm t)|^2dxdt\le 2a\int(|\dot{H}(x,0)|^2+|\dot{E}(x,0)|^2)dx$$

and we choose ε such that

$$\tag{4.12} 1-\frac{8\varepsilon a^2}{\alpha_*}\|M_0\|_\infty=\frac{1}{2}.$$

Finally, if we set

$$\tag{4.13} C_0=\frac{1}{2}+2a\varepsilon,\quad C_2=\frac{1}{2}+\frac{\|M_0\|_\infty}{4\varepsilon\alpha_*},\quad C_1=\varepsilon\frac{\|M_0\|_\infty}{\alpha_*},$$

we obtain

$$\tag{4.14} \left| \begin{aligned} &\frac{1}{2}\int(|\dot{H}(x,T)|^2+|\dot{E}(x,T)|^2)dx+\frac{1}{4}\int_{-a}^{a}\int_0^T\frac{\alpha}{|M|}|\ddot{M}|^2dxdt \\ &\le C_0\int(|\dot{H}(x,0)|^2+|\dot{E}(x,0)|^2)dx+C_1\int_{-a}^{a}\int_0^T\frac{\alpha}{|M|}|\dot{M}|^2dxdt \\ &\qquad +C_2\int_{-a}^{a}\int_0^T\frac{\alpha}{|M|}|H_T|^2|\dot{M}|^2dxdt. \end{aligned}\right.$$

From (2.20), we already know that

$$\int_{-a}^{a}\int_{0}^{T}\frac{\alpha}{|M|}|\dot{M}|^2 dxdt < +\infty$$

while, because of the regularity assumptions on the initial data, one can see that

$$\int |\dot{E}(x,0)|^2 dx \;=\; \int |\text{curl } H_0|^2 dx < +\infty. \tag{4.15}$$

Conversely,

$$\int |\dot{H}(x,0)|^2 dx \;\le\; 2\int |\text{curl } E_0|^2 dx \;+\; 2\int |\dot{M}(x,0)|^2 dx, \tag{4.16}$$

and from the LLG equation one deduces that

$$|\dot{M}(x,0)|^2 = \frac{1}{1+\alpha^2}|H_T(x,0)|^2|M(x,0)|^2.$$

Therefore, setting $H_T^0(x) = H_T(x,0)$, we have

$$\int |\dot{M}(x,0)|^2 dx \le \|M_0\|_{L^\infty}^2 \|H_T^0\|_{L^2}^2 \tag{4.17}$$

which is finite since, under the conditions on the data of the problem

$$H_T^0 = H_0 - KP(M_0) + H_s \;\in\; L^2, \tag{4.18}$$

we can conclude that

$$\int |\dot{H}(x,0)|^2 dx \le 2\|\text{curl } E_0\|_{L^2}^2 + 2\|M_0\|_{L^\infty}^2\|H_T^0\|_{L^2}^2. \tag{4.19}$$

Finally, if C denotes a constant which depends only on M_0, E_0, H_0, α_*, K, H_s, and a, we can write

$$\left|\begin{aligned} &\frac{1}{2}\int(|\dot{H}(x,T)|^2 + |\dot{E}(x,T)|^2)dx + \frac{1}{4}\int_0^T\int_{-a}^{a}\frac{\alpha}{|M|}|\ddot{M}|^2 dxdt \\ &\qquad\qquad \le C + C_2\int_0^T\int_{-a}^{a}\frac{\alpha}{|M|}|\dot{M}|^2|H_T|^2 dxdt. \end{aligned}\right. \tag{4.20}$$

Step 4: *A Gronwall-type estimate.* From the definition of H_T, we easily get the following bound for $|H_T|^2$:

$$|H_T|^2 \le 3(1+\|K\|_\infty^2)\left[|H_s|^2 + |M|^2 + |H|^2\right]$$

that implies, since $H_x = -M_x$,

$$|H_T|^2 \le 3(1+\|K\|_\infty^2)(|H_s|^2 + 2|M|^2) + 3(1+\|K\|_\infty^2)|H_\parallel|^2. \tag{4.21}$$

Using the interpolation inequality

$$\|H_\parallel\|_{L^\infty}^2 \;\le\; \|H_\parallel\|_{L^2}\left\|\frac{\partial H_\parallel}{\partial x}\right\|_{L^2},$$

we can write

$$\|H_T\|^2_{L^\infty} \;\le\; 3(1+\|K\|^2_\infty)(\|H_s\|^2_{L^\infty}+2\|M_0\|^2_{L^\infty}) + 3(1+\|K\|^2_\infty)\|H_\|\|_{L^2}\left\|\frac{\partial H_\|}{\partial x}\right\|_{L^2}. \tag{4.22}$$

As we already know from Theorem 2.4 that $\|H_\|\|_{L^2}$ remains bounded, from (4.20) and (4.22), we can write

$$\left|\begin{array}{l} \displaystyle\frac{1}{2}\int(|\dot H(x,T)|^2+|\dot E(x,T)|^2)dx + \frac{1}{4}\int_0^T\int_{-a}^a \frac{\alpha}{|M|}|\ddot M|^2 dxdt \\ \displaystyle\qquad\qquad\qquad \le C\left(1+\int_0^T\left(\int_{-a}^a \frac{\alpha}{|M|}|\dot M|^2dx\right)\left\|\frac{\partial H_\|}{\partial x}\right\|_{L^2}dt\right), \end{array}\right. \tag{4.23}$$

where C denotes a new positive constant depending only on the data of the problem. Finally, using Maxwell's equations, we observe that

$$\left\|\frac{\partial H_\|}{\partial x}\right\|_{L^2} = \left\|\dot E_\|\right\|_{L^2}. \tag{4.24}$$

Therefore, if we set

$$\left\{\begin{array}{rcl} G(t) &=& \displaystyle\int(|\dot H(x,t)|^2+|\dot E(x,t)|^2)dx, \\ m(t) &=& \displaystyle\int_{-a}^a \frac{\alpha}{|M|}|\dot M|^2dx, \end{array}\right. \tag{4.25}$$

we deduce from (4.23), using that $\int_0^\infty\int_{-a}^a \frac{\alpha}{|M|}|\dot M|^2dxdt < +\infty$, the inequality ($C$ denoting another constant)

$$G(t) \;\le\; C\left(1+\int_0^t m(s)G(s)^{\frac{1}{2}}ds\right)$$

with $m \in L^1(0,+\infty)$. Therefore, using an appropriate generalization of Gronwall's lemma, we get

$$G(t) \le \left(C^{\frac{1}{2}} + C\int_0^t m(s)ds\right)^2. \tag{4.26}$$

As $m \in L^1(0,+\infty)$, this shows that $G(t)$ remains bounded in time. Moreover, from (4.23) we also deduce

$$\int_0^T\int_{-a}^a \frac{\alpha}{|M|}|\ddot M|^2dxdt \le C\left(1+\int_0^T G(s)^{\frac{1}{2}}m(s)ds\right). \tag{4.27}$$

Setting $G^* = \sup_{t\ge 0} G(t)$, we thus obtain

$$\int_0^T\int_{-a}^a \frac{\alpha}{|M|}|\ddot M|^2dxdt \le C\left(1+(G^*)^{\frac{1}{2}}\|m\|_{L^1}\right) \tag{4.28}$$

which concludes the proof of the lemma. □

Proof of Theorem A$'$. First, note that the inequalities

$$\left|\begin{array}{l} \displaystyle\int_{-a}^{a} |\dot M(x,t)|^2 dx \leq \frac{\|M_0\|_{L^\infty}}{\alpha_*} \int_{-a}^{a} \frac{\alpha}{|M|} |\dot M|^2 dx, \\ \displaystyle\int_{-a}^{a} |\ddot M(x,t)|^2 dx \leq \frac{\|M_0\|_{L^\infty}}{\alpha_*} \int_{-a}^{a} \frac{\alpha}{|M|} |\ddot M|^2 dx, \end{array}\right.$$

joined to the results of Theorem A and Lemma 4.1, imply

$$\int_0^\infty \int_{-a}^{a} \left(|\dot M|^2 + |\ddot M|^2 \right) dxdt < +\infty;$$

that is to say,

$$\dot M \in H^1\left(\mathbb{R}^+, L^2(-a,a)\right), \tag{4.29}$$

which implies

$$\lim_{t\to+\infty} \int_{-a}^{a} |\dot M(x,t)|^2 dx = 0. \tag{4.30}$$

Now, let us return to the formula

$$|\dot F(x,t)| \leq \frac{\sqrt{2}}{2} \int_{\Gamma_a(x,t)} |\ddot M(y,s)| dy + \frac{1}{2} \left\{ |\dot F^0(x+t)| + |\dot F^0(x-t)| \right\}.$$

By the same manipulations as in the proof of Theorem A (cf. the method we used to obtain (3.9)), we get

$$\left|\begin{array}{l} \displaystyle\int_{-R}^{R} |\dot F(x,t)|^2 dx \leq \frac{4a\|M_0\|_\infty}{\alpha_*} \int_{t-(R+a)}^{t} \int_{-a}^{a} \frac{\alpha}{|M|} |\ddot M(y,s)|^2 dyds \\ \displaystyle\qquad\qquad + \int_{-R-t}^{R-t} |\dot F^0(x)| dx + \int_{-R+t}^{R+t} |\dot F^0(x)| dx. \end{array}\right. \tag{4.31}$$

Since $\int_{-\infty}^{+\infty} |\dot F^0(x)|^2 dx < +\infty$ and $\int_0^\infty \int_{-a}^{a} \frac{\alpha}{|M|} |\ddot M|^2 dyds < +\infty$ (cf. Lemma 2.1), (4.31) implies that

$$\lim_{R\to+\infty} \int_{-R}^{R} \left(\left|\dot E_\parallel\right|^2 + \left|\dot H_\parallel\right|^2 \right) dx = 0. \tag{4.32}$$

Now, using Maxwell's equations we have

$$\left|\begin{array}{l} \displaystyle\int_{-R}^{R} \left|\frac{\partial E_\parallel}{\partial x}\right|^2 dx = \int_{-R}^{R} \left|\dot H_\parallel\right|^2 dx, \\ \displaystyle\int_{-R}^{R} \left|\frac{\partial H_\parallel}{\partial x}\right|^2 dx \leq 2 \int_{-R}^{R} \left|\dot E_\parallel\right|^2 dx + 2 \int_{-a}^{a} |\dot M|^2 dx \end{array}\right. \tag{4.33}$$

which, taking into account (4.30) and (4.32), lead to (2.26). Thanks to Theorem A, (2.27) is then a consequence of (2.26), since $H^1(-R,R) \subset L^\infty(-R,R)$. This concludes the proof of the theorem. □

5. Transitions to stationary states in the LLG equation. This section must be seen as an introduction to the second part of the paper, which is devoted to long-time asymptotics of the magnetization M.

5.1. Stationary states in the LLG equation. We consider the solutions to the following nonlinear evolution equation that we shall call the unperturbed LLG equation at point x:

$$\left\{ \begin{array}{l} \dot{M}(x,t) = \tilde{H}_T(x, M(x,t)) \times M(x,t) + \dfrac{\alpha}{|M(x,t)|} M(x,t) \times \dot{M}(x,t), , \\ M(x, t=0) = M_0(x), \end{array} \right. \tag{5.1}$$

where $\tilde{H}_T(x, M(x,t))$ has been defined by (2.28).

Remark 5.1. One can easily check that the first equation of (5.1) can be rewritten as

$$\dot{M}(x,t) = L(x, M(x,t)), \tag{5.2}$$

where we have defined

$$L(x,M) \;=\; \frac{1}{1+\alpha^2} \left[\tilde{H}_T(x,M) \times M + \frac{\alpha}{|M|} M \times \Big(\tilde{H}_T(x,M) \times M \Big) \right]. \tag{5.3}$$

We now introduce the set $\mathcal{S}(x)$ of the stationary states for (5.1) as

$$\mathcal{S}(x) \;=\; \Big\{ M_0 \in \mathbb{R}^3 / \dot{M}(x,t) \;=\; 0 \quad \forall t > 0 \Big\}. \tag{5.4}$$

The main property of this set is that its intersection with any sphere $\Sigma(R)$ is a discrete set. (See Figure 5.1).

Theorem 5.2. *Let $R > 0$. Under assumption ($\mathcal{H}_x$) (see section* 2.2*), the intersection of $\mathcal{S}(x)$ with the sphere $\Sigma(R)$ with center at the origin and of radius R is a set which contains at least two elements and at most six.*

This result is of interest because (see (2.9))

$$|\dot{M}| \;=\; 0 \quad \Leftrightarrow \quad |\tilde{H}_T(x,M) \times M| \;=\; 0.$$

This shows that, for the set $\mathcal{Z}(x, M_0)$ defined in section2.2,

$$\mathcal{Z}(x, M_0) \;=\; \mathcal{S}(x) \cap \Sigma(|M_0(x)|),$$

and we have the following corollary.

Corollary 5.3. *The set $\mathcal{Z}(x, M_0)$ is a discrete set which contains at least two elements and at most six.*

Proof. We give a geometrical proof of Theorem 5.2: we characterize $\mathcal{S}(x)$ in every possible case. Toward this end, let us consider

$$\tilde{H}_T(x,M) \times M \;=\; 0. \tag{5.5}$$

This equality leads to two different problems depending on whether or not H_s is equal to 0.

Case 1. $H_s = 0$. If M belongs to $\mathcal{S}(x)$, we deduce from (5.5) that

$$\exists\, \lambda \in \mathbb{R}, \quad K\,(p \cdot M)p \,-\, (e_x \cdot M)e_x \;=\; \lambda M, \tag{5.6}$$

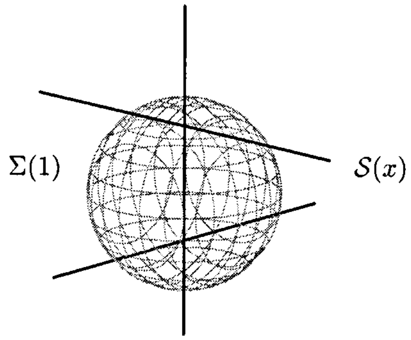


FIG. 5.1. *The unit sphere and the set $\mathcal{S}(x)$ with six intersection points ($H_s = 0.5\,e_z$, $K = 0.7$, $p = e_y$).*

which corresponds to a simple eigenvalue problem. Assumption $(\mathcal{H}_x)$ ensures that p and e_x are not collinear and that $K \neq 0$. Thus the operator defined by $M \mapsto K(p \cdot M)p - (e_x \cdot M)e_x$ is a real symmetric operator A; in the basis $(e_x, p, e_x \times p)$, we have

$$A \;=\; \begin{bmatrix} -1 & -(e_x \cdot p) & 0 \\ K(e_x \cdot p) & K & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The operator A has three distinct real eigenvalues: one is zero, and the other ones have opposite signs (as $K\left[(e_x \cdot p)^2 - 1\right] < 0$). Thus there exist three different eigendirections, (e_i) and $\mathcal{S}(x)$ is made up of three lines passing through the center of the $\Sigma(R)$. There are exactly six stationary states such that $|M_0| = R$, namely $M_0 = Re_i$, $i = 1, \ldots, 3$.

Case 2. $H_s \neq 0$. Let us denote $u = e_x$, $v = p$, and $w = H_s/|H_s|$. We shall distinguish three cases:

- *Case* 2.1 (*general case, see Figure* 5.2): $(u, v, w) \neq 0$.

  In the basis $\{u, v, w\}$, we denote M by $(x, y, z)^t$. Equality (5.5) then reads

  $$(\,|H_s|\,w \;+\; Ky\,v \;-\; x\,u\,) \times (\,x\,u \;+\; y\,v \;+\; z\,w\,) \;=\; 0,$$

  which yields

  $$|H_s|x\; w{\times}u + |H_s|y\; w{\times}v + Kxy\; v{\times}u + Kyz\; v{\times}w - xy\; u{\times}v - xz\; u{\times}w \;=\; 0.$$

  As $\{u \times v, u \times w, v \times w\}$ is also a basis of $\mathbb{R}^3$, we deduce that

  $$\text{(5.7)} \qquad \left\{ \begin{array}{rcl} x\,(|H_s| + z) & = & 0, \\ y\,(|H_s| - Kz) & = & 0, \\ x\,y\,(K+1) & = & 0. \end{array} \right.$$

  This shows that $\mathcal{S}(x)$ is made up of two or three lines: $d_1 = \{x = 0, y = 0\}$, $d_2 = \{y = 0, z = -|H_s|\}$ and, if $K \neq 0$, $d_3 = \{x = 0, z = |H_s|/K\}$. We deduce that the intersection between $\mathcal{S}(x)$ and a sphere of radius R has at least two points (because the line d_1 passes through the point $(0, 0, 0)$) and at most six points. All the values between 2 and 6 are possible, depending on R and on the distance of the lines d_2 and d_3 from the point $(0, 0, 0)$.

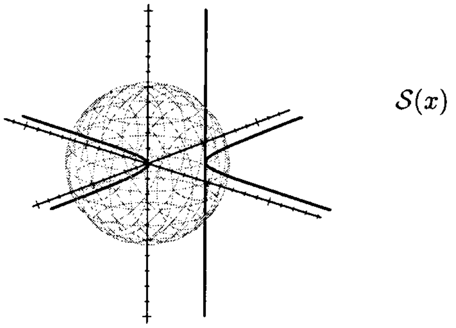


FIG. 5.2. *The unit sphere and the set $\mathcal{S}(x)$ with six intersection points ($H_s = 0.5\,(e_x + p)$, $K = 0.5$, $p = e_y$).*

- *Case* 2.2: $(u, v, w) = 0$, $(u, v, u \times v) \neq 0$.
  In the basis $\{u, v, u \times v\}$, we denote M by $(x, y, z)^t$ and H_s by $(h_u, h_v, 0)^t$. The computations now lead to

$$(5.8)\qquad \left\{\begin{array}{ccc} x\,(h_v + Ky) \;-\; y\,(h_u + x) & = & 0, \\ z\,(h_v + Ky) & = & 0, \\ z\,(h_u - x) & = & 0. \end{array}\right.$$

  In this case, the set $\mathcal{S}(x)$ is made up of a hyperbola ($\{z = 0, (K+1)xy + h_v x - h_u y = 0\}$) and, if $K \neq 0$, a line ($\{x = h_u, y = -h_v/K\}$). As one of the branches of the hyperbola passes through the point $(0, 0, 0)$, the conclusions are exactly the same as in the previous case.
- *Case* 2.3: u, v, and w are collinear.
  In the canonical basis $\{e_x, e_y, e_z\}$, we have $p = e_x$, $H_s = |H_s| e_x$, and $M = (x, y, z)^t$. The computations now lead to

$$(5.9)\qquad \left\{\begin{array}{ccc} y\,(|H_s| + Ky - x) & = & 0, \\ z\,(|H_s| + Ky - x) & = & 0. \end{array}\right.$$

  As

$$(\mathcal{H}_x) \quad \Leftrightarrow \quad (|H_s| + Ky - x) \;\neq\; 0,$$

  the set $\mathcal{S}(x)$ reduces in this case to the line $\{y = 0, z = 0\}$, and there are always exactly two intersection points. □

5.2. Free transitions to stationary states. In this section we still consider the unperturbed case (i.e., $H_\parallel = 0$). It is indeed easy to show in this case the convergence of $M(x, t)$ to some element of $\mathcal{Z}(x, M_0)$. It is also possible to see which of these positions are stable and which are not (see Remark 5.6).

THEOREM 5.4. *Let $\mathcal{Z}(x, M_0)$ be defined as in section* 2.2. *The solution to system* (5.1) *satisfies*

$$\exists\, M_\infty(x) \in \mathcal{Z}(x, M_0) \quad \textit{such that} \quad \lim_{t\to\infty} M(x, t) \;=\; M_\infty(x).$$

Proof. Note that

$$(5.10)\qquad \tilde{H}_T(x, M(x,t)) \;=\; H_s(x) + K(p \cdot M(x,t))p - (e_x \cdot M(x,t))e_x$$

can also be defined as

$$\tilde{H}_T(x, M(x,t)) \;=\; (\nabla_M V)\,(x, M(x,t)), \tag{5.11}$$

where we define

$$V(x,M) \;=\; H_s(x)\cdot M(x,t) - \frac{1}{2}(e_x \cdot M(x,t))^2 - \frac{1}{2}|P(M)|^2 \quad \forall x \in \mathbb{R},\ \forall M \in \mathbb{R}^3.$$

For this reason, we have

$$\tilde{H}_T(x, M(x,t)) \cdot \dot{M}(x,t) \;=\; \frac{d}{dt}V(x, M(x,t)). \tag{5.12}$$

Additionally, we deduce from (5.2) that

$$\tilde{H}_T(x, M(x,t)) \cdot \dot{M}(x,t) \;=\; \frac{\alpha}{1+\alpha^2}\,\frac{1}{|M|}\left|\tilde{H}_T(x, M(x,t)) \times M(x,t)\right|^2. \tag{5.13}$$

Thus

$$\frac{d}{dt}V(x, M(x,t)) \;=\; g(x, M(x,t)), \tag{5.14}$$

where

$$g(x,M) \;=\; \frac{\alpha}{1+\alpha^2}\,\frac{1}{|M|}\left|\tilde{H}_T(x,M) \times M\right|^2 \quad \forall x \in \mathbb{R},\ \forall M \in \mathbb{R}^3. \tag{5.15}$$

The main properties of this function g are, respectively,

(i) $g(x,M) \;\geq\; 0$;

(ii) $(\,g(x,M) = 0 \text{ and } |M| = |M_0(x)|\,) \;\Leftrightarrow\; (M \in \mathcal{Z}(x, M_0))\,$;

(iii) $g(x,M) \;=\; \frac{\alpha}{1+\alpha^2}\,\frac{1}{|M|}\,|\nabla_M V(x,M) \times M|^2$.

Therefore, (5.14) means that the function $M \to V(x,M)$ is a strict anti-Liapunov function for the system (5.1). Classical results on dynamic systems then ensure the convergence to these stationary states (see, for instance, [6]). □

Remark 5.5. Property (iii) of function g means, in particular, that the set of extremal points of $V(x,M)$ on the sphere is included in $\mathcal{Z}(x, M_0)$.

Moreover, in the general case ($(e_x, p, H_s) \neq 0$), it can be shown that, if $M \in \mathcal{Z}(x, M_0)$ is neither a maximum nor a minimum for V, then it is a saddlepoint for V (it cannot be a local extremum).

As an illustration of these results and Theorem 5.2, we represent below two trajectories of the vector $M(t)$ on the sphere of radius $|M_0| = 1$ (see Figure 5.3). These trajectories have been computed numerically and correspond to the following data: ($H_s = 2\,e_z$, $K = 0$) and ($H_s = 0.5\,e_z$, $K = 0.7$, $p = e_y$), respectively. One can check that the first case corresponds to two stationary states on the sphere while the second one corresponds to six states (all are indicated by bold arrows). The initial position, indicated by a bold dot, is the same in both cases.

Remark 5.6. Although it is not the main objective of this paper, one could complete the analysis by results on the stability of the stationary points. Let us first recall the following definition.

Definition 5.7. *A stationary state M_∞ is stable for the trajectory $M(x,t)$ if and only if*

$$\exists\, \mathcal{V}(M_\infty)\ \forall M_0 \in \mathcal{V}(M_\infty), \quad \lim_{t\to\infty} M(x,t) \;=\; M_\infty, \tag{5.16}$$

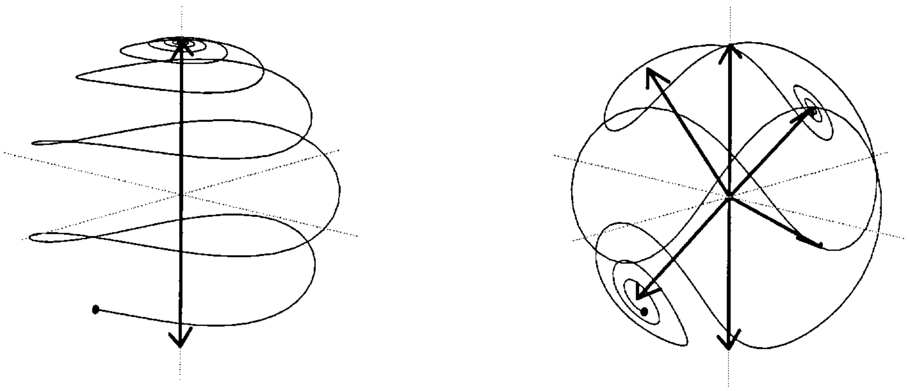


FIG. 5.3. *Trajectories of $M(t)$ with two and six stationary states.*

where $\mathcal{V}(M_\infty)$ is a neighborhood of M_∞ and $M(x,t)$ is the solution to (5.1) *associated with the initial data M_0. (Otherwise, M_∞ is said to be unstable.)*

Two results can now be pointed out:

1. In the general case $((e_x, p, H_s) \neq 0)$, one can show that only the points $M_\infty \in \mathcal{Z}(x, M_0)$, such that

$$V(M_\infty) \;=\; \max_{M \in S} V,$$

are stable stationary states, while the others are unstable.

2. This is not necessarily true when $(e_x, p, H_s) = 0$.

6. Proof of Theorem B: Attraction of M for weak solutions. Using the notation of section 5, one can check that the LLG equation of (1.5) can be rewritten (see Remark 5.1) in the form

$$\dot{M}(x,t) \;=\; L(x, M(x,t)) + R(x,t), \tag{6.1}$$

where

$$R(x,t) = \frac{1}{1+\alpha^2}\left[H_\parallel(x,t) \times M(x,t) + \frac{\alpha}{|M|} M(x,t) \times \big(H_\parallel(x,t) \times M(x,t)\big)\right]. \tag{6.2}$$

By the definition of V and g (see section 5, proof of Theorem 5.4), one easily shows that

$$\frac{d}{dt} V(x, M(x,t)) = g(x, M(x,t)) + r(x,t), \;\; t > 0, \tag{6.3}$$

where

$$r(x,t) \;=\; -\tilde{H}_T(x, M(x,t)) \cdot R(x,t) + \frac{\alpha}{1+\alpha^2} \frac{1}{|M|} \left|H_\parallel(x,t) \times M(x,t)\right|^2. \tag{6.4}$$

As a consequence of Theorem A and (2.6), we see that $R(x,M)$ satisfies

$$\int_0^\infty |R(x, M(x,t), H_\parallel(x,t))|^2 dt \;<\; \infty \text{ for a.e. } x \in \mathbb{R}. \tag{6.5}$$

and thus

$$\int_0^\infty |r(x, M(x,t), H_\|(x,t))|dt \; < \; \infty \text{ for a.e. } x \in \mathbb{R}. \tag{6.6}$$

Let us fix an $x \in \mathbb{R}$ such that (6.1)–(6.6) hold. For this reason we shall write $M(t)$, $R(t)$, and $r(t)$ instead of $M(x,t)$, $R(x,t)$, and $r(x,t)$, respectively, and similarly $g(M)$ and $V(M)$ instead of $g(x,M)$ and $V(x,M)$. By Theorem 2.8, we know that all these functions are, almost everywhere in x, continuous functions of time. In the following, we shall consider such an x. We shall also write $\mathcal{Z}$ instead of $\mathcal{Z}(x, M_0)$ and denote by S the sphere of radius $|M_0(x)|$.

The first step of the proof is the convergence of $V(M(t))$. Let $V(\mathcal{Z}) = \{V_0, V_1, \ldots, V_N\}$ with $V_0 < V_1 < \cdots < V_N$ (In particular, by Remark 5.5 V_0 is the minimum of V on S and V_N is the maximum). Let us introduce

$$V^+ = \limsup_{t\to\infty} V(M(t)) \in]V_i, V_{i+1}] \quad \text{for some } i. \tag{6.7}$$

Lemma 6.1. *Let V_{i+1} be defined by* (6.7). *Then we necessarily have*

$$V(M(t)) \to V_{i+1} \textit{ as } t \to \infty. \tag{6.8}$$

Proof. We prove (6.8) by contradiction. Let us assume to the contrary that

$$V^- \; = \; \liminf_{t\to\infty} \; V(M(t)) \; < \; V_{i+1}.$$

Using also (6.7), we deduce that there exists $\varepsilon > 0$ and $V^0 \in \mathbb{R}$ such that

$$\max(V^-, V_i + \varepsilon) < V^0 < \min(V^+, V_{i+1} - \varepsilon). \tag{6.9}$$

Therefore, by continuity of $V(M(t))$ there exists a sequence $t_k \to \infty$ such that

$$V(M(t_k)) = V^0. \tag{6.10}$$

If we prove that, for sufficiently large T and $t > T$,

$$V(M(t)) \; \geq \; V_{i+1} - \varepsilon, \tag{6.11}$$

we will get a contradiction to (6.10) since $t_k \to \infty$ and $V^0 < V_{i+1} - \varepsilon$. To prove (6.11), first note that the properties of g imply the existence of some $\delta > 0$ such that

$$V(M) \in \left[V_i + \frac{\varepsilon}{2}, V_{i+1} - \frac{\varepsilon}{2}\right] \; \Rightarrow \; g(M) \geq \delta. \tag{6.12}$$

Let us introduce for each k the set

$$\mathcal{E}_k = \left\{ t > t_k : \; V(M(\tau)) \in \left[V_i + \frac{\varepsilon}{2}, V_{i+1} - \frac{\varepsilon}{2}\right] \text{ for } t_k < \tau < t \right\}, \tag{6.13}$$

It is easy to see that by construction $\mathcal{E}_k$ is connected and that, taking into account (6.9) and (6.10) as well as the continuity of $V(M(t))$, it is not empty. One can also prove that this set is bounded. Indeed, let $\tau \in \mathcal{E}_k$; integrating (6.3) between t_k and τ, we get

$$V(M(\tau)) - V(M(t_k)) \; = \; \int_{t_k}^{\tau} g(M(s))ds + \int_{t_k}^{\tau} r(s)ds. \tag{6.14}$$

By definition of $\mathcal{E}_k$, we have

$$\forall\, t_k < s < \tau, \quad V(M(s)) \in \left[V_i + \frac{\varepsilon}{2}, V_{i+1} - \frac{\varepsilon}{2}\right].$$

Therefore, using (6.14) and (6.12), we get

$$V(M(\tau)) - V(M(t_k)) \;\geq\; \delta(\tau - t_k) - I_k, \tag{6.15}$$

and thus,

$$\tau \;\leq\; t_k + \frac{1}{\delta}\left(\max_{M\in S} V(M) - \min_{M\in S} V(M) + I_k\right), \tag{6.16}$$

where

$$I_k \;=\; \int_{t_k}^{\infty} |r(s)| ds. \tag{6.17}$$

Therefore $\mathcal{E}_k$ is bounded by

$$t_k + \frac{1}{\delta}\left(\max_{M\in S} V(M) - \min_{M\in S} V(M) + I_k\right). \tag{6.18}$$

Now, let us introduce

$$\tau_k = \sup \mathcal{E}_k < +\infty. \tag{6.19}$$

By continuity, $V(M(\tau_k))$ is equal to $V_i + \frac{\varepsilon}{2}$ or $V_{i+1} - \frac{\varepsilon}{2}$. Let us show that $V(M(\tau_k)) = V_{i+1} - \frac{\varepsilon}{2}$ at least for k large enough. Indeed, from (6.14), we deduce in particular that (g is positive and we take $\tau = \tau_k$)

$$V(M(\tau_k)) - V(M(t_k)) \;\geq\; -I_k. \tag{6.20}$$

I_k tends to 0 when k tends to $+\infty$ because of (6.6). Thus, there exists $\bar{k} = \bar{k}(\varepsilon)$ such that, for $k \geq \bar{k}$,

$$V(M(\tau_k)) - V(M(t_k)) \;\geq\; -\frac{\varepsilon}{2} \text{ for } t_k < \tau < \tau_k. \tag{6.21}$$

Therefore

$$V(M(\tau_k)) \geq V(M(t_k)) - \frac{\varepsilon}{2} = V^0 - \frac{\varepsilon}{2} > V_i + \frac{\varepsilon}{2}.$$

This proves that (see Figure 6.1)

$$k \geq \bar{k} \Rightarrow V(M(\tau_k)) = V_{i+1} - \frac{\varepsilon}{2}.$$

Now, we introduce the set

$$\mathcal{F}_k \;=\; \left\{t \geq \tau_k \,:\, V(M(\tau)) \geq V_{i+1} - \frac{\varepsilon}{2} \text{ for } \tau_k \leq \tau \leq t\right\}. \tag{6.22}$$

Equations (6.9) and (6.10) ensure that $\mathcal{F}_k$ is bounded for every k. Let us set

$$\theta_k = \max \mathcal{F}_k. \tag{6.23}$$

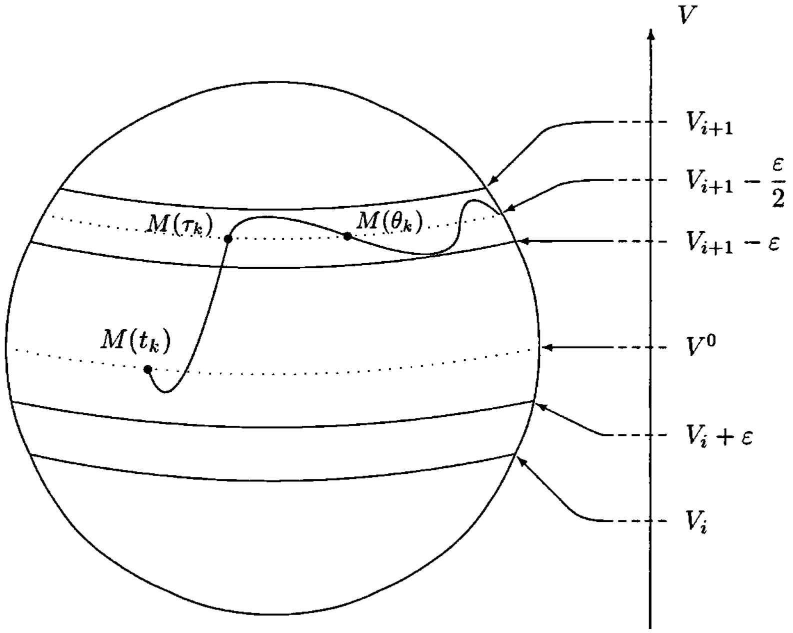


FIG. 6.1. *Behavior of a Liapunov function.*

By definition of $\mathcal{F}_k$ and continuity of $V(M(t))$, we have

$$V(M(\theta_k)) = V_{i+1} - \frac{\varepsilon}{2}. \tag{6.24}$$

Integrating (6.3) between θ_k and $t > \theta_k$ and using once again the positivity of g, we get, for $k \geq \overline{k}$,

$$\forall\, t > \theta_k \quad V(M(k)) - V_{i+1} + \frac{\varepsilon}{2} \geq -I_k \geq -\frac{\varepsilon}{2} \tag{6.25}$$

which proves (6.11) with $T = \theta_{\overline{k}}$. □

Now to prove Theorem B, let us assume by contradiction that the conclusion is not true. This means that there exists $\rho > 0$ and a strictly increasing sequence $t_k \to \infty$ such that (see Figure 6.2)

$$\operatorname{dist}(M(t_k), \mathcal{Z}) \;\geq\; \rho. \tag{6.26}$$

We are going to prove that this contradicts Lemma 6.1.

First, using the properties of g, we know that there exists $\delta > 0$ such that

$$\operatorname{dist}(M(t_k), \mathcal{Z}) \;\geq\; \frac{\rho}{2} \;\Rightarrow\; g(M) \geq \delta. \tag{6.27}$$

By continuity of $M(t)$ and by (6.26), we know that the set

$$\mathcal{G}_k \;=\; \left\{ t > t_k \,/\, \tau \in]t_k, t[\Rightarrow \operatorname{dist}(M(\tau), \mathcal{Z}) \;\geq\; \frac{\rho}{2} \right\}$$

is not empty. Let us check that, for k large enough,

$$]t_k, t_k + T] \subset \mathcal{G}_k \;\text{ where }\; T \;=\; C\rho, C > 0. \tag{6.28}$$

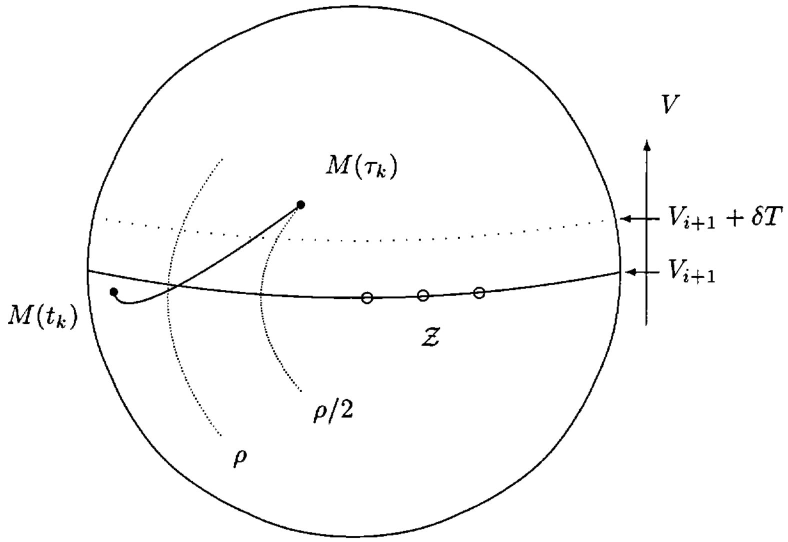


FIG. 6.2

Let $\tau_k = \sup \mathcal{G}_k$. As the case $\tau_k = +\infty$ is obvious ($]t_k, t_k+T] \subset]t_k, +\infty[$), let us assume that $\tau_k < +\infty$. By continuity, $\operatorname{dist}(M(\tau_k), \mathcal{Z}) = \rho/2$, and using (6.26),

$$|M(t_k) - M(\tau_k)| \geq \frac{\rho}{2}. \tag{6.29}$$

Conversely, integrating (6.1) between t_k and τ_k and using (6.5) and the Cauchy–Schwarz inequality, we get

$$|M(\tau_k) - M(t_k)| = \left| \int_{t_k}^{\tau_k} L(M(s))ds + \int_{t_k}^{\tau_k} R(s)ds \right| \leq \omega(\tau - t_k) + J_k(\tau_k - t_k)^{\frac{1}{2}}, \tag{6.30}$$

where $\omega > 0$, and

$$J_k^2 = \int_{t_k}^{\infty} |R(s)|^2 ds. \tag{6.31}$$

Regrouping (6.29) and (6.30), we get, using Young's inequality,

$$\frac{\rho}{2} \leq \omega(\tau_k - t_k) + J_k(\tau_k - t_k)^{\frac{1}{2}} \leq 2\omega(\tau_k - t_k) + \frac{J_k^2}{2\omega}. \tag{6.32}$$

As J_k^2 tends to 0 when $k \to \infty$, we have, for $k \geq \tilde{k}$, $J_k^2/2\omega < \rho/4$. Therefore,

$$k \geq \tilde{k} \Rightarrow \frac{\rho}{4} \leq 2\omega(\tau_k - t_k) \Rightarrow (\tau_k - t_k) \geq CT. \tag{6.33}$$

Now, integrating (6.3) between t_k and $t_k + T$, we get

$$V(M(t_k+T)) - V(M(t_k)) = \int_{t_k}^{t_k+T} g(M(s))ds + \int_{t_k}^{t_k+T} r(s)ds. \tag{6.34}$$

By (6.26)–(6.28), $g(M(s)) \geq \delta$ for $s \in [t_k, t_k + T]$ and $k \geq \tilde{k}$. Therefore,

$$V(M(t_k + T)) - V(M(t_k)) \ \geq \ \delta T \ - \ I_k, \tag{6.35}$$

where $I_k = \int_{t_k}^{+\infty} |r(s)| ds$ tends to 0 when $k \to +\infty$. Therefore, for k large enough,

$$V(M(t_k + T)) - V(M(t_k)) \ \geq \ \frac{\delta T}{2} \tag{6.36}$$

which of course contradicts Lemma 6.1. □

7. On the attraction of M for strong solutions. In this last section, the assumptions are those of Theorem A′. First, note that the proof of Corollary 2.12 is obvious: indeed, the proof given in the previous section for weak solutions also applies to strong solutions and ensures the convergence of $M(x, t)$ for every $x \in \mathbb{R}$.

One could think that the uniform convergence to 0 of the transverse field $H_{\|}(x, t)$ would yield that the convergence of the magnetization distribution $M(x, t)$ to $M_\infty(x, t)$ is itself uniform. In fact, such a result is not obvious at all, and may not be true. More precisely, we are going to prove, with the help of a suitable counterexample, that it is not possible to prove the uniform convergence of $M(x, t)$ with the only assumption that the convergence of $H_{\|}(x, t)$ is uniform.

To construct this counterexample, we denote by S the unit sphere, set $M_\infty = e_z = (0, 0, 1)^t$, and $M_{-\infty} = -M_\infty$. We shall need two simple lemmas which apply to the evolution equation

$$\begin{cases} \dot{M}(t) = \tilde{H}_T(M(t)) \times M(t) + \dfrac{\alpha}{|M(t)|} M(t) \times \dot{M}(t), \\ M(t = 0) = M_0 \in S, \end{cases} \tag{7.1}$$

where $\tilde{H}_T(M(t)) = -(M(t) \cdot e_x)e_x + 2e_z + H_{\|}(t)$, which corresponds to $H_s = 2e_z$ and $K = 0$. In such a case, one easily verifies that

$$\mathcal{Z} \ = \ \{ M_\infty, M_{-\infty} \} .$$

LEMMA 7.1. *Assuming that $H_{\|}(t) = 0$, for any solution $M(t)$ to system* (7.1), *and for all $M_0 \in S$, we have*

$$(M_0 \neq M_{-\infty}) \Rightarrow \Big(V(M_0) \ > \ V_{\min} = V(M_{-\infty}) \text{ and } \lim_{t \to \infty} M(t) \ = \ M_\infty \Big) , \tag{7.2}$$

where

$$\forall M \in S, \quad V(M) \ = \ 2M \cdot e_z - \frac{1}{2} |M \cdot e_x|^2.$$

Remark 7.2. This result means that M_∞ can be a limit state if and only if $M_0 = M_{-\infty}$, in which case the solution is stationary.

Proof. It is easy to check that

$$\max_{M \in S} V(M) \ = \ V_{\max} \ = \ V(M_\infty),$$

and

$$\min_{M \in S} V(M) \ = \ V_{\min} \ = \ (V(M_{-\infty})).$$

This is true as soon as $|H_s| > 1$ (see section 4). Moreover, for the same reason, V admits no other critical point on S. The lemma is then a direct consequence of the fact that V is a strict anti-Liapunov function for the system (7.1). □

LEMMA 7.3. *Let $\varepsilon \in]0,1]$, and let us assume that $H_{\|}(t) = \varepsilon e_y$ and $M_0 = M_{-\infty}$. Then, for all $T > 0$, the solution to system* (7.1) *on* $[0,T]$ *is such that*

$$V(M(T)) \; < \; V_{\min} \; = \; V(M_{-\infty})$$

where V is as defined in Lemma 7.1.

Proof. In this case, the total magnetic field is

$$\tilde{H}_T(M) = -(M(t) \cdot e_x)e_x + 2e_z + \varepsilon e_y.$$

The associated strict anti-Liapunov function is

$$\forall M \in S, \; V'(M) \; = \; -\frac{1}{2}|M \cdot e_x|^2 + 2M \cdot e_z + \varepsilon M \cdot e_y, \tag{7.3}$$

which is such that

$$\min_{M \in S} V'(M) \; = \; V'_{\min} \; = \; V'(M_{-\infty}) \; = \; V(M_{-\infty}),$$

where V has been defined in Lemma 7.1, and

$$\max_{M \in S} V'(M) \; = \; V'_{\max} \; = \; V'(M_{\infty}) \; = \; V(M_{\infty}).$$

Thus,

$$V'(M(T)) \; > \; V'_{\min} \; = \; V(M_{\infty}). \tag{7.4}$$

Let us define

$$\mathcal{C}' \; = \; \{M \in S \setminus V'(M) \; = \; V'(M(T))\} \tag{7.5}$$

and

$$V_T \; = \; \min_{M \in \mathcal{C}'} V(M). \tag{7.6}$$

As $V'_{\min} = V_{\min}$ is reached in M_{∞} only, we see that

$$(\, V_T = V_{\min} \,) \implies (\, V'(M(T)) = V'_{\min} \,)\,, \tag{7.7}$$

which is not true; whence the result, since

$$V'(M(T)) \; \geq \; V_T \; > \; V_{\min}. \qquad \square \tag{7.8}$$

We can now state the following theorem.

THEOREM 7.4. *Let $\Omega = [0,1]$ be a ferromagnetic layer defined by the initial distribution*

$$M_0(x) \; = \; M_{-\infty} \quad \forall x \in [0,1],$$

$H_s = 2e_z$, and $K = 0$. Let us consider a transverse magnetic field defined by

$$\left| \begin{array}{lcll} H_{\|}(x,t) & = & 0 & \forall x \notin [0,1]\,, \\ & = & 0 & \forall t \notin \left[\dfrac{1}{x}, \dfrac{2}{x}\right], \\ & = & x & \forall t \in \left[\dfrac{1}{x}, \dfrac{2}{x}\right]. \end{array} \right.$$

Then, $H_{\|}(x,t)$ *converges uniformly to* 0 *when* t *goes to* $+\infty$; *however,*

(i) $\lim_{t\to\infty} M(0,t) \;=\; M_{-\infty}$*;*

(ii) $\forall x \in]0,1]$*,* $\lim_{t\to\infty} M(x,t) \;=\; M_{\infty}$*;*

(iii) $\forall \varepsilon \in [0,2[\; \forall T > 0$*,* $\exists\, x \in]0,1]$*,* $\exists\, t > T$ *such that*

$$\left| M(x,t) - \lim_{t\to\infty} M(x,t) \right| \;=\; |\, M(x,t) - M_\infty \,| \;=\; 2 \;> \varepsilon.$$

Proof. First of all, it is easy to see that

$$\sup_{x\in\,]0,1]} H_{\|}(x,t) \;=\; \frac{2}{t} \tag{7.9}$$

which ensures that the convergence of $H_{\|}$ to 0 in time is uniform in space.

Concerning the three assertions of the theorem, result (i) is clear since $H_{\|}(0,t) = 0$ for all $t \geq 0$. For (ii) and (iii), let us consider $\varepsilon \in [0,2[$ and $T > 0$. We choose $x \in]0,1]$ such that $1/x > T$. Then, on the one hand,

$$\forall t \in [0,T], \quad H_{\|}(x,t) = 0 \;\text{ and } M(x,t) \;=\; M_{-\infty}. \tag{7.10}$$

On the other hand, by Lemma 7.3,

$$V\left(M\left(x, t = \frac{2}{x}\right)\right) \;>\; V_{\min} \tag{7.11}$$

and thus, by Lemma 7.1,

$$\lim_{t\to\infty} M(x,t) \;=\; M_\infty \qquad \square \tag{7.12}$$

Remark 7.5. The previous example is only a counterexample to the fact that the uniform convergence of M could be proven by using uniform convergence of $H_{\|}$. It is not a counterexample to the uniform convergence of M: indeed, it is not a solution of the coupled Maxwell–LLG system.

Remark 7.6. One might also believe that $M(t)$ always converges to a stable stationary state. This is not so obvious, and may not be true. In fact, one can show that a "perturbation" $H_{\|}(x,t)$—that is to say, a function C^1 in time vanishing to 0 with $t \to \infty$— can lead the magnetization M from a stable position to an unstable one. Let us consider

$$\left\{ \begin{array}{l} \dot\mu(t) = -\,\mu(t) \times (H_T(\mu(t)) \times \mu(t)), \\ \mu(t=0) = \mu_0 \in S, \end{array} \right. \tag{7.13}$$

where $H_T(\mu) = -(\mu \cdot e_x)e_x + 2e_z$. The solution to this problem is such that

$$\forall \mu_0 \neq M_\infty, \quad \lim_{t\to\infty} \mu(t) \;=\; M_{-\infty}, \tag{7.14}$$

because $V(\mu)$ is a strict anti-Liapunov function for (7.13). Let us now define the function $h(t)$ as

$$h(t) \;=\; \frac{1}{1+\alpha^2} H_T(\mu(t)) \times \mu(t) - \frac{1+\alpha+\alpha^2}{1+\alpha^2}\mu(t) \times (H_T(\mu(t)) \times \mu(t)). \tag{7.15}$$

This function is such that

(i) $h(t) \in C^1(\mathbb{R})$ because $\mu(t) \in C^1(\mathbb{R})$;

(ii) $h(t) \to 0$ when $t \to \infty$ because $\mu(t) \times H_T(\mu(t)) \to 0$ when $t \to \infty$;

(iii) additionally, computations lead to

$$(H_T(\mu(t)) + h(t)) \times \mu(t) + \alpha\mu(t) \times [(H_T(\mu(t)) + h(t)) \times \mu(t)] = \\ -\ \mu(t) \times (H_T(\mu(t)) \times \mu(t)).$$

In other words, function $\mu(t)$, the solution to problem (7.13), is also the solution to problem (7.1) with the perturbation $H_{\|}(t) = h(t)$.

REFERENCES

[1] W. B. Brown, *Micromagnetics*, Interscience Tracts of Physics, Wiley Intescience, New York, 1963.

[2] B. Lax and K. J. Button, *Microwave Ferrites and Ferrimagnetics*, McGraw-Hill, New York, 1962.

[3] W. L. Miranker and B. E. Willner, *Global analysis of magnetic domains*, Quart. Appl. Math., 37 (1979/80), pp. 219–238.

[4] A. Visintin, *On Landau–Lifschitz' equations for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.

[5] J. L. Joly, G. Métivier, and J. Rausch, *On Landau–Lifschitz' Equations for Ferromagnetism*, in preparation.

[6] A. Haraux, *Nonlinear Evolution Equations–Global Behavior of Solutions*, Springer-Verlag, Berlin, 1981.

[7] P. Joly and O. Vacus, *Maxwell's Equations in a* 1*D Ferromagnetic Medium: Existence and Uniqueness of Strong Solutions*, Rapport de Recherche INRIA 3052, France, 1997.

[8] P. Joly and O. Vacus, *Propagation d'ondes en milieu ferromagnétique* 1*D: Existence et unicité de solutions faibles*, Note soumise au Compte Rendu de l'Académie des Sciences, 1996.

[9] P. D. Lax, C. S. Morawetz, and R. S. Phillips, *Exponential decay of solutions of the wave equation in the exterior of a star-shaped obstacle*, Comm. Pure Appl. Math., 16 (1963), pp. 477–486.

[10] C. S. Morawetz, *The decay of solutions of the exterior initial-boundary value problem for the wave equation*, Comm. Pure Appl. Math., 14 (1961), pp. 561–568.

[11] C. S. Morawetz and W. A. Strauss, *Decay and scattering of solutions of a nonlinear relativistic wave equation*, Comm. Pure Appl. Math., 25 (1972), pp. 1–31.

[12] A. Komech, *On transitions to stationary states in some infinite dimensional Hamiltonian systems*, Dokl. Math., 53 (1996), pp. 208–210.

[13] A. I. Komech, *On the stabilization of string-oscillators interaction*, Russian J. Math. Phys., 3 (1995), pp. 227–248.

[14] A. I. Komech and H. Spohn, *Soliton-like asymptotics for a classical particle interacting with a scalar wave field*, Nonlinear Anal., 33 (1998), pp. 13–24.

[15] A. I. Komech, H. Spohn, and M. Kunze, *Long-time asymptotics for a classical particle interacting with a scalar wave field*, Comm. Partial Differential Equations, 22 (1997), pp. 307–335.

[16] A. I. Komech, *On stabilization of string-nonlinear oscillator interaction*, J. Math. Anal. Appl., 196 (1995), pp. 384–409.

[17] A. I. Komech and B. R. Vainberg, *On asymptotic stability of stationary solutions to nonlinear wave and Klein-Gordon equations*, Arch. Rational Mech. Anal., 134 (1996), pp. 227–248.

[18] P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 375–385

ON THE SOLUTION OF A LINEAR HOMOGENEOUS DIFFERENCE EQUATION WITH VARIABLE COEFFICIENTS[*]

RANJAN K. MALLIK[†]

Abstract. A closed form solution of a linear homogenous difference equation of order N with variable coefficients is presented for $N \geq 3$. From it, the solution for the special case of an equation with constant coefficients is also obtained.

Key words. Nth order linear homogeneous difference equation, linear recurrence, variable coefficients, closed form solution

AMS subject classification. 39A10

PII. S0036141097329640

1. Introduction.

Linear homogeneous difference equations or linear homogeneous recurrences play a significant role in the mathematical models of systems which we come across in various areas of science and engineering. Closed form solutions of linear recurrences with constant coefficients are known, and the recurrences have a nice and complete theory. That is no longer the case when the coefficients vary with the index [1]. For equations with variable coefficients, the closed form solution of the first order equation is known [2, 3, 4]. However, for homogeneous equations of order N greater than one, if $N-1$ linearly independent solutions are known, then any other linearly independent solution can be obtained in closed form in terms of the known solutions using the Casoratian (Wronskian in discrete domain) [2, 3, 5, 6]. For a second order linear homogeneous difference equation with variable coefficients, the closed form solution in terms of coefficients has been presented in [7, 8]. The solution of the Nth order linear difference equation with variable coefficients has also been represented in terms of determinants of submatrices of a single solution matrix [9]. But, in the available open literature, there are no general closed form expressions in terms of coefficients for the complete solution of linear homogeneous difference equations with varying coefficients when the order is N, except for cases when the coefficients have some special properties [1]. This paper presents *a complete closed form solution in terms of coefficients* of a linear homogeneous difference equation of order N with variable coefficients when $N \geq 3$.

We begin with some definitions and examine some properties based on the definitions. Let $\mathbf{N}$ denote the set of natural numbers. For any given $N \in \mathbf{N}$, $N \geq 3$, and for all $n \in \mathbf{N}$, define a function $f : \mathbf{N} \to \mathbf{N}$ as

$$f(n) \triangleq \begin{cases} (N-1)\lfloor \frac{n}{N} \rfloor + 1 & \text{if } n \text{ is divisible by } N, \\ (N-1)\lfloor \frac{n}{N} \rfloor + N - 1 & \text{if } n \text{ is not divisible by } N, \end{cases} \tag{1.1}$$

where $\lfloor \frac{n}{N} \rfloor$ is the largest natural number which is less than or equal to $\frac{n}{N}$. For all ascending q-tuples $(k_1, \ldots, k_q)$ of natural numbers, where $q \geq 2$ and $k_1 < \cdots < k_q$,

[*] Received by the editors November 5, 1997; accepted for publication (in revised form) December 1, 1998; published electronically January 5, 2000.

http://www.siam.org/journals/sima/31-2/32964.html

[†] Department of Electrical Engineering, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi 110016, India (rkmallik@ee.iitd.ernet.in).

define a function $\mathbf{g}_q : \mathbf{N}^q \to \mathbf{N}^q$ as

$$
\begin{array}{l}
\mathbf{g}_q(k_1, \ldots, k_q) \triangleq (l_1, \ldots, l_q)\,, \\
\text{such that} \\
l_1 = 2, \ \text{ and} \\
\text{for } m = 2, \ldots, q, \\
l_m = \begin{cases} 1 + l_{m-1} & \text{if } k_m - k_{m-1} = 1, \\ 2 & \text{if } k_m - k_{m-1} \geq 2. \end{cases}
\end{array}
\tag{1.2}
$$

For example, $\mathbf{g}_2(3,4) = (2,3)$, $\mathbf{g}_2(3,5) = \mathbf{g}_2(3,6) = (2,2)$, $\mathbf{g}_3(3,4,5) = (2,3,4)$, $\mathbf{g}_3(4,5,7) = (2,3,2)$, $\mathbf{g}_4(3,4,6,7) = (2,3,2,3)$, $\mathbf{g}_4(3,5,6,7) = (2,2,3,4)$. Thus the function $\mathbf{g}_q$ keeps track of the strings of consecutive numbers in $(k_1, \ldots, k_q)$.

Let a set $S_q(L,U)$, where $q, L, U \in \mathbf{N}$, be defined as the set of all q-tuples with elements from $\{L, L+1, \ldots, U\}$ arranged in ascending order in which no N consecutive elements are present, that is,

$$
\begin{array}{lll}
(1.3\text{a}) & S_q(L,U) \triangleq \{L, L+1, \ldots, U\} & \text{if } \ U \geq L \ \text{ and } \ q = 1 \\
(1.3\text{b}) & \quad\quad\quad \triangleq \left\{ {\scriptstyle (k_1,\ldots,k_q)\,:\, L \leq k_1 < \cdots < k_q \leq U} \right\} & \text{if } \ U \geq L + q - 1, \ \ q = 2, \ldots, N-1 \\
& \quad\quad\quad \triangleq \left\{ {\scriptstyle (k_1,\ldots,k_q)\,:\, L \leq k_1 < \cdots < k_q \leq U;\ \ k_m - k_{m-N+1} \geq N \ \text{ for } \ m = N, \ldots, q} \right\} & \\
(1.3\text{c}) & \quad\quad\quad\quad \text{if } \ U \geq L + N \ \text{ and } \ N \leq q \leq f(U-L) & \\
(1.3\text{d}) & \quad\quad\quad \triangleq \emptyset \quad \text{otherwise.} &
\end{array}
$$

Also, let a set $T_q(L,U)$ be defined as

$$
T_q(L,U) \triangleq \left\{ {\scriptstyle ([k_1,l_1],\ldots,[k_q,l_q])\,:\,(k_1,\ldots,k_q) \in S_q(L,U);\ \ (l_1,\ldots,l_q) \in \mathbf{g}_q(k_1,\ldots,k_q)} \right\}, \tag{1.4}
$$

where functions f and $\mathbf{g}_q$ are defined in (1.1) and (1.2), respectively.

Proposition 1. *For $U \geq L+1$, $2 \leq q \leq f(U-L+N)$,*

$$
\begin{aligned}
S_q(L, U+N) = S_q(L, U+N-1) & \\
& \cup \left\{ {\scriptstyle (k_1,\ldots,k_q)\,:\,(k_1,\ldots,k_{q-1}) \in S_{q-1}(L,U+N-2);\ \ k_q = U+N} \right\} \\
& \cup \left\{ {\scriptstyle (k_1,\ldots,k_q)\,:\,(k_1,\ldots,k_{q-2}) \in S_{q-2}(L,U+N-3);\ \ k_{q-1} = U+N-1,\ \ k_q = U+N} \right\} \\
& \cup \cdots \\
& \cup \left\{ {\scriptstyle (k_1,\ldots,k_q)\,:\,(k_1,\ldots,k_{q-N+1}) \in S_{q-N+1}(L,U);\ \ k_{q-N+2} = U+2,\ldots,k_q = U+N} \right\} \\
= S_q(L, U+N-1) \cup & \\
& \left[\bigcup_{r=1}^{N-1} \left\{ {\scriptstyle (k_1,\ldots,k_q)\,:\,(k_1,\ldots,k_{q-r}) \in S_{q-r}(L,U+N-r-1);\ \ k_{q-p} = U+N-p,\ p=0,\ldots,r-1} \right\} \right].
\end{aligned}
\tag{1.5}
$$

Proof. From the definition of $S_q(L,U)$, $S_q(L,U+N)$ can be expressed as a disjoint union of $\mathcal{S}_r$, $r = 0, \ldots, N-1$, where $\mathcal{S}_r$ is the set of q-tuples $(k_1, \ldots, k_q) \in S_q(L, U+N)$ such that $k_{q-r} \leq U + N - r - 1$, and

$$
k_{q-r+1} = U + N - r + 1,\ k_{q-r+2} = U + N - r + 2,\ \ldots,\ k_q = U + N.
$$

Note that

$$\mathcal{S}_r = \left\{ (k_1,\ldots,k_{q-r},U+N-r+1,\ldots,U+N) : (k_1,\ldots,k_{q-r}) \in S_{q-r}(L,U+N-r-1) \right\},$$

so the proposition immediately follows. □

PROPOSITION 2. *For $U \geq L+1$, $2 \leq q \leq f(U-L+N)$,*

$$\begin{aligned} T_q(L,U+N) = T_q(L,U+N-1) \cup \\ \left[\bigcup_{r=1}^{N-1} \left\{ ([k_1,l_1],\ldots,[k_q,l_q]) : ([k_1,l_1],\ldots,[k_{q-r},l_{q-r}]) \in T_{q-r}(L,U+N-r-1); \right. \right. \\ \left. \left. [k_{q-p},l_{q-p}] = [U+N-p,r+1-p],\ p=0,\ldots,r-1 \right\} \right]. \end{aligned} \tag{1.6}$$

Proof. From Proposition 1 and the definition of $T_q(L,U)$ in (1.4), we obtain (1.6). □

2. Solution of the difference equation. Consider the Nth order linear homogeneous difference equation ($N \geq 3$)

$$y_{n+N} = a_{n,1}\,y_{n+N-1} + a_{n,2}\,y_{n+N-2} + a_{n,3}\,y_{n+N-3} + \cdots + a_{n,N}\,y_n\,, \quad n \geq 1, \tag{2.1}$$

with integral index n, variable complex coefficients $a_{n,1},\ldots,a_{n,N}$, and complex initial values $y_1,\ldots,y_N$.

Since our basic objective is to obtain the solution in terms of relations between the coefficient indices directly or indirectly, we approach the problem by *transforming* the original difference equation into another difference equation with new coefficients expressed in terms of the original coefficients. The transformed difference equation is then solved in closed form in terms of the new coefficients.

We begin by defining a quantity $\gamma_{i,j}$ for $i \geq 2$, $j \geq 2$ as

$$\gamma_{i,j} \triangleq \begin{cases} \frac{a_{i,j}}{a_{i-1,j-1}a_{i,1}} & \text{if } j = 2,\ldots,N, \\ 0 & \text{if } j \geq N+1. \end{cases} \tag{2.2}$$

It is clear that the solution of difference equation (2.1) with initial values $y_1,\ldots,y_N$ can also be written as

$$\begin{aligned} y_i &= c_{i-N,1}\,y_N + c_{i-N,2}\,y_{N-1} + c_{i-N,3}\,y_{N-2} + \cdots + c_{i-N,N}\,y_1 \\ &= \sum_{j=1}^{N} c_{i-N,j}\,y_{N+1-j}\,, \quad i \geq 1, \end{aligned} \tag{2.3}$$

with

$$\begin{bmatrix} c_{0,1} & c_{0,2} & \cdots & c_{0,N} \\ c_{-1,1} & c_{-1,2} & \cdots & c_{-1,N} \\ \vdots & \vdots & & \vdots \\ c_{-(N-1),1} & c_{-(N-1),2} & \cdots & c_{-(N-1),N} \end{bmatrix} = \mathbf{I}_N\,, \tag{2.4}$$

$\mathbf{I}_N$ being the $N \times N$ identity matrix, where

$$c_{i,j} = a_{i,1}c_{i-1,j} + a_{i,2}c_{i-2,j} + a_{i,3}c_{i-3,j} + \cdots + a_{i,N}c_{i-N,j}\,, \quad i \geq 1,\ j = 1, \ldots, N, \tag{2.5}$$

i.e., $y_i = c_{i-N,j}$, $j = 1, \ldots, N$ are N solutions of difference equation (2.1) for $i \geq 1$.

Equation (2.5) can be rewritten as

$$\begin{aligned} c_{1,j} &= a_{1,j} && \text{for } \ j = 1, \ldots, N, \\ c_{i,j} &= \sum_{l=1}^{i-1} a_{i,l}c_{i-l,j} \ + \ a_{i,i+j-1} && \text{for } \ 2 \leq i \leq N+1-j, \ \ j = 1, \ldots, N-1 \\ &= \sum_{l=1}^{i-1} a_{i,l}c_{i-l,j} && \text{for } \ N+2-j \leq i \leq N, \ \ j = 2, \ldots, N \\ &= \sum_{l=1}^{N} a_{i,l}c_{i-l,j} && \text{for } \ \ i \geq N+1, \ \ j = 1, \ldots, N. \end{aligned} \tag{2.6}$$

Next, we define, for $j = 1, \ldots, N$,

$$\begin{aligned} d_{1,j} &\triangleq \tfrac{c_{1,j}}{a_{1,j}}\,, \\ d_{i,j} &\triangleq \tfrac{c_{i,j}}{a_{1,j}a_{2,1}\ldots a_{i,1}} \quad \text{for } \ i \geq 2. \end{aligned} \tag{2.7}$$

Using (2.7) and the definition of $\gamma_{i,j}$ in (2.2), we obtain, from (2.6),

$$\begin{aligned} d_{1,j} &= 1\,, \\ d_{2,j} &= d_{1,j} + \gamma_{2,j+1} \\ &= 1 + \gamma_{2,j+1}\,, \\ d_{i,j} &= d_{i-1,j} \ + \ \sum_{l=2}^{i-1} d_{i-l,j}\left[\prod_{p=2}^{l} \gamma_{i-l+p,p}\right] \ + \ \prod_{p=2}^{i} \gamma_{p,j+p-1} \quad \text{for } \ 3 \leq i \leq N \\ &= d_{i-1,j} \ + \ \sum_{l=2}^{N} d_{i-l,j}\left[\prod_{p=2}^{l} \gamma_{i-l+p,p}\right] \quad \text{for } \ i \geq N+1, \end{aligned} \tag{2.8}$$

where $j = 1, \ldots, N$.

It is clear from (2.8) that, for $j = 1, \ldots, N$, $i \geq 1$, the quantity $d_{i,j}$ can be expressed as

$$d_{i,j} = \phi_1(i) \ + \ \sum_{m=2}^{N} \phi_m(i)\left[\prod_{p=2}^{m} \gamma_{p,j+p-1}\right]\,, \tag{2.9}$$

where

$$\phi_1(1) = 1, \quad \phi_2(1) = \cdots = \phi_N(1) = 0, \\ \phi_1(2) = \phi_2(2) = 1, \quad \phi_3(2) = \cdots = \phi_N(2) = 0, \tag{2.10a}$$

$$\phi_m(i) = \phi_m(i-1) \ + \sum_{l=2}^{\min(i-1,N)} \phi_m(i-l) \left[\prod_{p=2}^{l} \gamma_{i-l+p,p}\right] \quad \text{for } \ i \geq 3, \tag{2.10b}$$

where $m = 1, \dots, N$. The *transformed difference equation* is given by (2.10b). Equation (2.10) also implies that

$$\begin{aligned}&\phi_1(1) = \phi_1(2) = 1, \quad \phi_1(3) = 1 + \gamma_{3,2}, \\ &\phi_m(i) = 0, \ \ 1 \leq i \leq m-1, \quad \phi_m(m) = \phi_m(m+1) = 1, \quad \phi_m(m+2) = 1 + \gamma_{m+2,2} \\ &\text{for } \ m = 2, \dots, N.\end{aligned} \tag{2.11}$$

The closed form solution of the transformed difference equation can be expressed in terms of $\gamma_{i,j}$s with their indices related by the set $T_q(L,U)$ defined in (1.4), as is given by the following proposition.

PROPOSITION 3. *For $m = 1, \dots, N$, $i \geq m+3$,*

$$\phi_m(i) \ = \ 1 \ + \sum_{k=m+2}^{i} \gamma_{k,2} \ + \sum_{q=2}^{q_{max}(i-m)} \ \sum_{([k_1,l_1],\dots,[k_q,l_q]) \in T_q(m+2,i)} (\gamma_{k_1,l_1} \cdots \gamma_{k_q,l_q}), \tag{2.12}$$

where f is defined in (1.1), *$T_q(L,U)$ in* (1.4), *$\gamma_{i,j}$ in* (2.2), *and*

$$q_{max}(i-m) = \begin{cases} i-m-1 & \textit{if} \ \ 3 \leq i-m \leq N-1, \\ f(i-m-2) & \textit{if} \ \ i-m \geq N. \end{cases} \tag{2.13}$$

Proof. From (2.10b) and (2.11), we obtain, for $m = 1, \dots, N$ (substituting $i = m+n$),

$$\phi_m(m+n) = \phi_m(m+n-1) \ + \sum_{r=2}^{\min(n,N)} \phi_m(m+n-r) \left[\prod_{p=2}^{r} \gamma_{m+n-r+p,p}\right] \\ \text{for } \ n \geq 3. \tag{2.14}$$

Using (2.10b) and (2.11), we get

$$\phi_m(m+3) = 1 + (\gamma_{m+2,2} + \gamma_{m+3,2}) + \gamma_{m+2,2}\gamma_{m+3,3}\,, \tag{2.15}$$

which satisfies (2.12) for all $N \geq 3$. Note that $q_{\max}(i-m) = 2$ when $i - m = 3$.

For $n \geq 4$, let (2.12) be valid for $\phi_m(m+n-1), \phi_m(m+n-2), \dots, \phi_m(m+3)$.

Then, applying (2.14), we obtain

$$\begin{aligned}\phi_m(m+n) \;=\; 1 \;&+\; \sum_{k=2}^{n-1}\gamma_{m+k,2} \\ &+ \sum_{q=2}^{q_{\max}(n-1)} \sum_{([k_1,l_1],\ldots,[k_q,l_q])\in T_q(2,n-1)} (\gamma_{m+k_1,l_1}\cdots\gamma_{m+k_q,l_q}) \\ &+ \sum_{r=2}^{\min(n,N)} (\gamma_{m+n-r+2,2}\cdots\gamma_{m+n,r}) \\ &+ \sum_{r=2}^{\min(n,N)} \sum_{k=2}^{n-r}\gamma_{m+k,2}(\gamma_{m+n-r+2,2}\cdots\gamma_{m+n,r}) \\ &+ \sum_{r=2}^{\min(n,N)} \sum_{q=r+1}^{q_{\max}(n-r)+r-1} \\ &\times \sum_{([k_1,l_1],\ldots,[k_{q-r+1},l_{q-r+1}])\in T_{q-r+1}(2,n-r)} (\gamma_{m+k_1,l_1}\cdots\gamma_{m+k_{q-r+1},l_{q-r+1}})(\gamma_{m+n-r+2,2}\cdots\gamma_{m+n,r})\,.\end{aligned}\tag{2.16}$$

Substituting $U = n - N$, $L = 2$ in (1.6), we get, for $2 \leq q \leq f(n-2)$,

$$\begin{aligned}&T_q(2,n) = T_q(2,n-1) \cup \\ &\left[\bigcup_{r=2}^{N} \{([k_1,l_1],\ldots,[k_q,l_q]) : ([k_1,l_1],\ldots,[k_{q-r+1},l_{q-r+1}])\in T_{q-r+1}(2,n-r); \right. \\ &\left. [k_{q-p},l_{q-p}]=[n-p,r-p],\ p=0,\ldots,r-2\}\right].\end{aligned}\tag{2.17}$$

Combining the q-product terms in (2.16) using (2.17) gives (2.12) when $i = m + n$. By mathematical induction, (2.12) is valid for all $i \geq m + 3$. □

From (2.3), (2.7), (2.9), and (2.2), we conclude that the *closed form solution* (in terms of the variable coefficients) of difference equation (2.1) with initial values $y_1, \ldots, y_N$, $N \geq 3$, is given by

$$y_{i+N} = c_{i,1}\,y_N + c_{i,2}\,y_{N-1} + c_{i,3}\,y_{N-2} + \cdots + c_{i,N}\,y_1 = \sum_{j=1}^{N} c_{i,j}\,y_{N+1-j}\,, \quad i \geq 1,\tag{2.18}$$

where

$$\begin{aligned}c_{1,j} &= a_{1,j}d_{1,j} = a_{1,j}\,, \quad j = 1,\ldots,N, \\ c_{i,j} &= a_{1,j}a_{2,1}\ldots a_{i,1}d_{i,j}\,, \quad j = 1,\ldots,N,\;\; i \geq 2,\end{aligned}\tag{2.19}$$

such that, for $j = 1, \ldots, N$, $i \geq 1$,

$$d_{i,j} = \phi_1(i) \;+\; \sum_{m=2}^{N} \phi_m(i)\left[\prod_{p=2}^{m}\gamma_{p,j+p-1}\right],\tag{2.20}$$

where $\gamma_{i,j}$ is defined in (2.2), and $\phi_m(i)$, $m = 1, \ldots, N$, are given by (2.11) and Proposition 3.

3. An example. Consider the case when $N = 3$ in difference equation (2.1). The initial values are y_1, y_2, y_3. Suppose we want to obtain an expression for y_6 in terms of the coefficients $a_{i,j}$, $j = 1, 2, 3$, $i = 1, 2, 3$. To start with, we find $\gamma_{i,j}$s and $\phi_m(i)$s. Equation (2.2) gives

$$\begin{aligned} \gamma_{2,2} &= \frac{a_{2,2}}{a_{1,1}a_{2,1}}\,, \quad \gamma_{3,2} = \frac{a_{3,2}}{a_{2,1}a_{3,1}}\,, \\ \gamma_{2,3} &= \frac{a_{2,3}}{a_{1,2}a_{2,1}}\,, \quad \gamma_{3,3} = \frac{a_{3,3}}{a_{2,2}a_{3,1}}\,, \end{aligned} \tag{3.1}$$

and, from (2.11), we get

$$\begin{array}{lll} \phi_1(1) = 1, & \phi_2(1) = 0, & \phi_3(1) = 0, \\ \phi_1(2) = 1, & \phi_2(2) = 1, & \phi_3(2) = 0, \\ \phi_1(3) = 1 + \gamma_{3,2}, & \phi_2(3) = 1, & \phi_3(3) = 1. \end{array} \tag{3.2}$$

Next, from (2.20) and (3.2), we obtain

$$\begin{aligned} d_{3,1} &= \phi_1(3) + \phi_2(3)\gamma_{2,2} + \phi_3(3)\gamma_{2,2}\gamma_{3,3} \\ &= 1 + \gamma_{3,2} + \gamma_{2,2} + \gamma_{2,2}\gamma_{3,3}\,, \\ d_{3,2} &= \phi_1(3) + \phi_2(3)\gamma_{2,3} \\ &= 1 + \gamma_{3,2} + \gamma_{2,3}\,, \\ d_{3,3} &= \phi_1(3) \\ &= 1 + \gamma_{3,2}\,. \end{aligned} \tag{3.3}$$

In addition, (2.19), (3.3), and (3.1) give

$$\begin{aligned} c_{3,1} &= a_{1,1}a_{2,1}a_{3,1}d_{3,1} = a_{1,1}a_{2,1}a_{3,1} + a_{1,1}a_{3,2} + a_{3,1}a_{2,2} + a_{3,3}\,, \\ c_{3,2} &= a_{1,2}a_{2,1}a_{3,1}d_{3,2} = a_{1,2}a_{2,1}a_{3,1} + a_{1,2}a_{3,2} + a_{3,1}a_{2,3}\,, \\ c_{3,3} &= a_{1,3}a_{2,1}a_{3,1}d_{3,3} = a_{1,3}a_{2,1}a_{3,1} + a_{1,3}a_{3,2}\,. \end{aligned} \tag{3.4}$$

Finally, (2.18) implies that

$$y_6 = c_{3,1}\, y_3 + c_{3,2}\, y_2 + c_{3,3}\, y_1\,, \tag{3.5}$$

where $c_{3,1}$, $c_{3,2}$ and $c_{3,3}$ are given by (3.4). This is the desired result, and it agrees with what is obtained by direct substitution.

4. Difference equation with constant coefficients. Consider the difference equation (2.1) with initial values $y_1, \ldots, y_N$, in which $a_{n,j} = a_j$ for all $i \geq 1$, $j = 1, \ldots, N$, i.e.,

$$y_{n+N} = a_1\, y_{n+N-1} + a_2\, y_{n+N-2} + a_3\, y_{n+N-3} + \cdots + a_N\, y_n\,, \quad n \geq 1. \tag{4.1}$$

As in (2.2), define

$$\gamma_j \triangleq \begin{cases} \frac{a_j}{a_{j-1}a_1} & \text{if } j = 2, \ldots, N, \\ 0 & \text{if } j \geq N+1. \end{cases} \tag{4.2}$$

The solution of difference equation (4.1) with initial values $y_1, \ldots, y_N$ is then given by (see (2.18), (2.19), (2.20))

$$y_{i+N} = c_{i,1}\, y_N + c_{i,2}\, y_{N-1} + c_{i,3}\, y_{N-2} + \cdots + c_{i,N}\, y_1\,, \quad i \geq 1, \tag{4.3}$$

where

$$c_{i,j} = a_j (a_1)^{i-1} d_{i,j}\,, \quad j = 1, \ldots, N, \;\; i \geq 1, \tag{4.4}$$

such that, for $j = 1, \ldots, N$, $i \geq 1$,

$$d_{i,j} = \phi_1(i) \; + \; \sum_{m=2}^{N} \phi_m(i) \left[\prod_{p=2}^{m} \gamma_{j+p-1} \right], \tag{4.5}$$

where, from (2.11) and (2.10),

$$\begin{aligned}
&\phi_1(1) = \phi_1(2) = 1, \quad \phi_1(3) = 1 + \gamma_2, \\
&\phi_m(i) = 0, \;\; 1 \leq i \leq m-1, \quad \phi_m(m) = \phi_m(m+1) = 1, \quad \phi_m(m+2) = 1 + \gamma_2 \\
&\text{for } \; m = 2, \ldots, N,
\end{aligned} \tag{4.6a}$$

$$\phi_m(i) = \phi_m(i-1) \; + \sum_{l=2}^{\min(i-1,N)} \phi_m(i-l) \left[\prod_{p=2}^{l} \gamma_p \right] \quad \text{for } \; i \geq 3. \tag{4.6b}$$

It is clear from (4.6) that

$$\phi_m(i) = \phi_1(i - (m-1)) \quad \text{for } \; m = 1, \ldots, N. \tag{4.7}$$

The condition $\phi_m(0) = 0$ satisfies the relation (4.6b). Let the generating function $\Phi_m(u)$ of $\phi_m(i)$ be defined as

$$\Phi_m(u) \triangleq \sum_{i=0}^{\infty} \phi_m(i) u^i\,. \tag{4.8}$$

Then, from (4.6) and (4.7), we obtain

$$\Phi_m(u) = \frac{u^m}{1 - u - \sum_{l=2}^{N} (\gamma_2 \ldots \gamma_l) u^l}\,. \tag{4.9}$$

Equation (4.2) gives

$$\gamma_2 \cdots \gamma_l = \frac{a_l}{a_1^l}\,, \quad l = 2, \ldots, N. \tag{4.10}$$

Expressing $\Phi_m(u)$ as a formal power series and using the multinomial expansion formula for each term of the series, we get, from (4.9) and (4.10),

$$\Phi_m(u) = u^m \sum_{q=0}^{\infty} \sum_{\substack{(k_1, \ldots, k_N) \\ k_1, \ldots, k_N \geq 0 \\ k_1 + k_2 + \cdots + k_N = q}} \binom{q}{k_1, \ldots, k_N} \left[\prod_{l=1}^{N} a_l^{k_l} \right] \left(\frac{u}{a_1} \right)^{k_1 + 2k_2 + \cdots + N k_N}. \tag{4.11}$$

Substituting $i-m = k_1 + 2k_2 + \cdots + Nk_N$ in (4.11), and using (4.8), we obtain

$$\phi_m(i) = \frac{1}{a_1^{i-m}} \sum_{\substack{(k_1,\ldots,k_N) \\ k_1,\ldots,k_N \geq 0 \\ k_1+2k_2+\cdots+Nk_N=i-m}} \begin{pmatrix} k_1+k_2+\cdots+k_N \\ k_1,\ldots,k_N \end{pmatrix} \left[\prod_{l=1}^{N} a_l^{k_l}\right], \tag{4.12}$$

for $i \geq m$. Note that $\phi_m(i) = 0$ for $0 \leq i \leq m-1$.

Now, (4.2) gives

$$\prod_{p=2}^{m} \gamma_{j+p-1} = \begin{cases} \frac{a_{j+m-1}}{a_j a_1^{m-1}} & \text{if } j+m-1 = 2,\ldots,N, \\ 0 & \text{if } j+m-1 \geq N+1, \end{cases} \tag{4.13}$$

which, along with (4.4) and (4.5), implies

$$c_{i,j} = \sum_{m=1}^{N+1-j} a_{j+m-1} a_1^{i-m} \phi_m(i)\,. \tag{4.14}$$

Substituting (4.12) in (4.14), we get

$$c_{i,j} = \sum_{m=j}^{N} \left\{ \sum_{\substack{(k_1,\ldots,k_N) \\ k_1,\ldots,k_N \geq 0 \\ k_1+2k_2+\cdots+Nk_N=i-m+j-1}} \begin{pmatrix} k_1+k_2+\cdots+k_N \\ k_1,\ldots,k_N \end{pmatrix} \left[\prod_{l=1}^{N} a_l^{k_l}\right] \right\} a_m\,. \tag{4.15}$$

After replacing k_m+1 by k_m in (4.15), the equation can be rewritten as

$$c_{i,j} = \sum_{\substack{(k_1,\ldots,k_N) \\ k_1,\ldots,k_N \geq 0 \\ k_1+2k_2+\cdots+Nk_N=i+j-1}} \frac{k_j+k_{j+1}+\cdots+k_N}{k_1+k_2+\cdots+k_N} \begin{pmatrix} k_1+k_2+\cdots+k_N \\ k_1,\ldots,k_N \end{pmatrix} \left[\prod_{l=1}^{N} a_l^{k_l}\right], \tag{4.16}$$

for $i \geq 1$, $j = 1,\ldots,N$. Therefore the solution of difference equation (4.1) with initial values $y_1,\ldots,y_N$ is given by (4.3) and (4.16). This is a closed form solution in terms of the coefficients $a_1,\ldots,a_N$ and does not require evaluation of the roots of the characteristic polynomial.

If the difference equation (4.1) is written in vector form as

$$\begin{bmatrix} y_{n+N} \\ y_{n+N-1} \\ \vdots \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{N-1} & a_N \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{n+N-1} \\ y_{n+N-2} \\ \vdots \\ y_n \end{bmatrix}, \quad n \geq 1, \tag{4.17}$$

where the $N \times N$ matrix on the right hand side of (4.17) is the *companion matrix*, and the initial value vector is $[y_N, y_{N-1}, \ldots, y_1]^T$, then the solution of (4.17) can also be obtained in terms of the powers of the companion matrix, that is,

$$\begin{bmatrix} y_{i+N} \\ y_{i+N-1} \\ \vdots \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{N-1} & a_N \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}^i \begin{bmatrix} y_N \\ y_{N-1} \\ \vdots \\ y_1 \end{bmatrix}, \quad i \geq 1. \tag{4.18}$$

Comparing (4.18) with (4.3) and (2.4), we get

$$\begin{aligned} &\begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{N-1} & a_N \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}^i \\ &\qquad = \begin{bmatrix} c_{i,1} & c_{i,2} & \cdots & c_{i,N-1} & c_{i,N} \\ c_{i-1,1} & c_{i-1,2} & \cdots & c_{i-1,N-1} & c_{i-1,N} \\ \vdots & \vdots & & \vdots & \vdots \\ c_{i-N+1,1} & c_{i-N+1,2} & \cdots & c_{i-N+1,N-1} & c_{i-N+1,N} \end{bmatrix} \end{aligned} \tag{4.19}$$

for $i \geq 0$.

Equation (4.19) implies that (4.16) and (2.4) give the expressions for the entries of any nonnegative power of the companion matrix in terms of the coefficients $a_1, \ldots, a_N$. The result is consistent with the expression for the combinatorial power of the companion matrix presented in Theorem 3.1 of [10], which has been proved by using a digraph to represent a matrix.

5. Conclusion. The closed form solution of the Nth order difference equation ($N \geq 3$) presented here makes use of combinatorial properties of the indices of the coefficients in an indirect manner. By substituting $N = 3$ and forcing the coefficient $a_{n,N}$ of the last term of (2.1) to zero, we can obtain the solution of the second order equation. The solution for the special case of the equation with constant coefficients gives a method of obtaining the nonnegative powers of the companion matrix.

Acknowledgment. The author expresses his gratitude to the anonymous referees for their constructive comments.

REFERENCES

[1] A. M. Odlyzko, *Linear recurrences with varying coefficients*, in Handbook of Combinatorics, Vol. 2, R. L. Graham, M. Grotschel, and L. Lovasz, eds., Elsevier, Amsterdam, 1995, pp. 1135–1138.

[2] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1978.

[3] W. G. Kelley and A. C. Peterson, *Difference Equations: An Introduction with Applications*, Academic Press, San Diego, 1991.

[4] S. N. Elaydi, *An Introduction to Difference Equations*, Springer-Verlag, New York, 1996.

[5] S. J. Farlow, *An Introduction to Differential Equations and Their Applications*, McGraw–Hill, New York, 1994.

[6] S. L. Ross, *Differential Equations*, 3rd ed., John Wiley, New York, 1984.

[7] J. Popenda, *One expression for the solutions of second order difference equations*, Proc. Amer. Math. Soc., 100 (1987), pp. 87–93.

[8] R. K. Mallik, *On the solution of a second order linear homogeneous difference equation with variable coefficients*, J. Math. Anal. Appl., 215 (1997), pp. 32–47.

[9] R. K. Kittappa, *A representation of the solution of the nth order linear difference equation with variable coefficients*, Linear Algebra Appl., 193 (1993), pp. 211–222.

[10] W. Y. C. Chen and J. D. Louck, *The combinatorial power of the companion matrix*, Linear Algebra Appl., 232 (1996), pp. 261–278.

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 386–415

ON THE CONTINUATION OF AN INVARIANT TORUS IN A FAMILY WITH RAPID OSCILLATIONS[*]

CARMEN CHICONE[†] AND WEISHI LIU[†]

Dedicated to Professor Jack Hale on the occasion of his seventieth birthday.

Abstract. A persistence theorem for attracting invariant tori for systems subjected to rapidly oscillating perturbations is proved. The singular nature of these perturbations prevents the direct application of the standard persistence results for normally hyperbolic invariant manifolds. However, as is illustrated in this paper, the theory of normally hyperbolic invariant manifolds, *when combined with an appropriate continuation method,* does apply.

Key words. persistence, continuation, rapid oscillation, invariant torus

AMS subject classifications. 34C15, 58F30

PII. S0036141098338740

1. Introduction.

We will prove a persistence theorem for attracting invariant tori for systems subjected to rapidly oscillating perturbations. The singular nature of these perturbations prevents the direct application of the standard persistence results for normally hyperbolic invariant manifolds. However, as we will illustrate in this paper, the theory of normally hyperbolic invariant manifolds, *when combined with an appropriate continuation method,* does apply.

Systems with rapidly oscillating perturbations arise naturally when a priori stable systems are periodically forced. In fact, partial averaging (perhaps to some high order) at a resonant torus together with a rescaling to slow time produces a system with a rapidly oscillating perturbation. For example, systems of this type are obtained in the dissipative periodically forced oscillator models introduced in [2].

We will consider the existence of invariant tori for a smooth family of differential equations of the form

$$\dot{x} = f(x, \epsilon) + \epsilon g(x, t, \epsilon),$$

where g is a periodic function of the independent variable. For systems of this type, the behavior of the subsystem

$$\dot{x} = f(x, \epsilon)$$

at $\epsilon = 0$ is important. If this subsystem has a normally hyperbolic invariant manifold at $\epsilon = 0$, then the persistence of this manifold into the full system is a result of the general persistence theorems of Fenichel [3] and Hirsch, Pugh, and Shub [7]. However, in many applications, either there is an invariant manifold at $\epsilon = 0$ that is not normally hyperbolic or the system is singular at $\epsilon = 0$. For example, the first situation will arise when a periodically forced oscillator is averaged at a resonance. In these cases,

[*] Received by the editors May 14, 1998; accepted for publication (in revised form) May 18, 1999; published electronically January 5, 2000.

http://www.siam.org/journals/sima/31-2/33874.html

[†] Department of Mathematics, University of Missouri, Columbia, MO 65211 (carmen@chicone.math.missouri.edu, liu@math.missouri.edu). The research of the first author was supported by the National Science Foundation under grant DMS-9303767.

a change of coordinates and a rescaling of the independent variable often yields an equivalent family, for $\epsilon \neq 0$, of the form

$$y' = F(y) + \epsilon G(y, \tau/\epsilon, \epsilon) \tag{1.1}$$

with new independent variable τ. While the subsystem

$$y' = F(y) \tag{1.2}$$

of the family (1.1) in these new coordinates no longer depends on the perturbation parameter, and is therefore regular, the singular nature of the perturbation is reflected in system (1.1) by rapid oscillations of the perturbation term in the slow time.

If the subsystem (1.2) has a normally hyperbolic invariant torus, then a small C^1 perturbation will also have a normally hyperbolic invariant torus by the general persistence theorems mentioned above. However, in the full system (1.1), the perturbation is not defined at $\epsilon = 0$. Also, we note that the perturbation is not C^1 small. In fact, the partial derivative with respect to τ can be large relative to ϵ. Thus, the usual persistence theory does not apply directly. To overcome this difficulty we will use the idea introduced by Kopell [8] of embedding the system into an auxiliary family given by

$$y' = F(y) + \delta G(y, \tau/\epsilon, \epsilon). \tag{1.3}$$

If $\epsilon > 0$ is fixed and $\delta > 0$ is sufficiently small, then, by the usual persistence theory, the system (1.3) has a normally hyperbolic invariant manifold. We will show that if $\epsilon > 0$ is sufficiently small, then this normally hyperbolic invariant manifold can be continued in this family to $\delta = \epsilon$.

The plan of the paper is as follows. In section 2, a description of the origin of the model system that we will study is given. In section 3 we discuss some previous work by Kopell [8] on the continuation problem for system (1.3). A conceptual gap in this work will be described. Also, we will present the general method for continuation that is used in this paper. The statement of our main theorem on the existence of invariant tori is in section 4, and the remaining sections of the paper are devoted to its proof.

2. A periodically perturbed oscillator. In this section we will briefly describe the origin of the explicit perturbation problem that we will study; see [2] for more details.

Consider a periodically perturbed planar oscillator given by

$$\dot{u} = f(u) + \epsilon g(u, t), \tag{2.1}$$

where, for each $u \in \mathbb{R}^2$, the function $t \mapsto g(u, t)$ is $2\pi/\Omega$ periodic and ϵ is a small parameter. Let us assume that the unperturbed system

$$\dot{u} = f(u) \tag{2.2}$$

is Hamiltonian and it has a regular period annulus $\mathcal{A}$, that is, an annulus consisting entirely of periodic orbits such that the associated period function is regular. Also, for each point ζ in the domain of definition of the system (2.1), let $t \mapsto u(t, \zeta, \epsilon)$ denote the solution of (2.1) with the initial condition $u(0, \zeta, \epsilon) = \zeta$.

A periodic orbit Γ in $\mathcal{A}$ with period T is called resonant if there are relatively prime positive integers m and n such that

$$m\frac{2\pi}{\Omega} = nT. \tag{2.3}$$

If Γ is a resonant periodic orbit and $p \in \Gamma$, then the associated (subharmonic) Melnikov function is given by

$$M^{m:n}(\phi) := \int_0^{m2\pi/\Omega} f(u(t,p,0)) \wedge g(u(t,p,0), t-\phi)\, dt. \tag{2.4}$$

From a geometric point of view, the sign of the Melnikov function on a resonant orbit determines the "drift direction" for perturbed orbits. If, for example, the Melnikov function has a fixed sign, then perturbed orbits drift away from the vicinity of the resonant orbit in a direction determined by this sign; while if the Melnikov function has a simple zero, then there is a nearby perturbed periodic orbit; see [5], [9], [11].

If the Melnikov function vanishes identically on a resonant orbit, then a reasonable expectation is that the corresponding unperturbed torus in the phase cylinder, corresponding to the unperturbed resonant orbit, persists under the perturbation. If the perturbation is dissipative, then the perturbed invariant torus is an attractor. The presence of this attractor is often one of the dominant features of the global dynamics: perturbed orbits are entrained to this torus. Thus, the existence of invariant tori is an important consideration in the analysis of the global dynamics of the system. However, as we will see, the proof of the existence of an attracting invariant torus in this context requires additional hypotheses as well as a delicate perturbation analysis.

To study the dynamics of differential (2.1) near a resonant periodic orbit, it is convenient to consider the system in action angle coordinates. In fact, there is a smooth change of coordinates in a neighborhood of the resonant orbit such that the differential (2.1), in the new coordinates (I, ϑ), has the form

$$\dot{I} = \epsilon F(I,\vartheta,t), \quad \dot{\vartheta} = \omega(I) + \epsilon G(I,\vartheta,t), \tag{2.5}$$

where both F and G are 2π periodic in ϑ and $2\pi m/\Omega$ periodic in t. In these coordinates, the resonant orbit is given by $\{(I,\vartheta) : I = I_0\}$, where

$$m\frac{2\pi}{\Omega} = n\frac{2\pi}{\omega(I_0)}. \tag{2.6}$$

A "normal form" for system (2.5) with $\epsilon > 0$ at the resonant orbit is obtained by using the coordinate transformation

$$I = I_0 + \sqrt{\epsilon}\,\ell, \qquad \vartheta = \omega(I_0)t + \sigma,$$

followed by the Taylor expansion of the resulting vector field to third order in powers of $\sqrt{\epsilon}$. The transformed system

$$\begin{aligned}
\dot{\ell} =& \sqrt{\epsilon}\, F(I_0, \omega(I_0)t+\sigma, t) + \epsilon F_I(I_0, \omega(I_0)t+\sigma, t)\ell \\
&+ \epsilon^{3/2} F_{II}(I_0, \omega(I_0)t+\sigma, t)\ell^2 + O(\epsilon^2), \\
\dot{\sigma} =& \sqrt{\epsilon}\, \omega'(I_0)\ell + \epsilon\Big(G(I_0, \omega(I_0)t+\sigma, t) + \frac{1}{2}\omega''(I_0)\ell^2\Big) \\
&+ \epsilon^{3/2}\Big(G_I(I_0, \omega(I_0)t+\sigma, t)\ell + \frac{1}{6}\omega'''(I_0)\ell^3\Big) + O(\epsilon^2)
\end{aligned} \tag{2.7}$$

is in the "time periodic standard form," the correct form for averaging. Under the assumption that the Melnikov function vanishes on the resonant orbit, that is, the average of F in the new coordinates vanishes, there is an averaging transformation (slightly more general than the transformation used in [2] where the average of G is also assumed to vanish) such that the averaged system has the abstract form

$$\begin{aligned}\dot{\ell} &= \mu^2 p(\sigma)\ell + \mu^3\big(q(\sigma)\ell^2 + r(\sigma)\big) + \mu^4 \widehat{R}(\ell,\sigma,t,\mu),\\ \dot{\sigma} &= \mu\lambda\ell + \mu^2(\nu\ell^2 + g(\sigma)) + \mu^3\widehat{S}(\ell,\sigma,t,\mu),\end{aligned} \tag{2.8}$$

where p, q, r, and g are 2π periodic functions, λ, μ, and ν are real numbers, and both of the functions $\widehat{R}$ and $\widehat{S}$ are 2π periodic in σ and $2\pi/\Omega$ periodic in t. In fact, all of the functions appearing in the system (2.8) are identifiable in terms of the original vector field. Also, the new small parameter is defined by $\mu := \sqrt{\epsilon}$.

Let us rewrite system (2.8) as the autonomous system

$$\begin{aligned}\dot{\ell} &= \mu^2 p(\sigma)\ell + \mu^3\big(q(\sigma)\ell^2 + r(\sigma)\big) + \mu^4 \widehat{R}(\ell,\sigma,\varphi,\mu),\\ \dot{\sigma} &= \mu\lambda\ell + \mu^2(\nu\ell^2 + g(\sigma)) + \mu^3\widehat{S}(\ell,\sigma,\varphi,\mu),\\ \dot{\varphi} &= 1,\end{aligned} \tag{2.9}$$

where φ is a new angular variable modulo $2\pi m/\Omega$. Also, let us assume that the family is class C^∞. We will seek an invariant torus for system (2.9) as the graph of a periodic function $(\sigma,\varphi) \mapsto h(\sigma,\varphi)$; that is, h is 2π periodic in σ and $2\pi m/\Omega$ periodic in φ.

The Lyapunov–Perron method is used in [2] to prove the following theorem.

THEOREM 2.1. *Consider the differential* (2.9) *and define*

$$M := \min_{0\le\sigma\le 2\pi} |p(\sigma)| > 0.$$

If $g(\sigma) \equiv 0$, $\lambda \neq 0$,

$$5M > \operatorname{Lip}(p), \qquad M^2 \ge 6|\lambda|\|r\|_{0,1}, \tag{2.10}$$

and μ *is sufficiently small, then there is a periodic function* $h \in C^{0,1}$ *(supremum + Lipschitz norm) such that its graph* $\{(\ell,\sigma,t) : \ell = h(\sigma,t)\}$ *is an invariant torus for* (2.9).

Here, M is a measure of the minimum "normal contraction rate" and the inequality $M^2 \ge 6|\lambda|\|r\|_{0,1}$ is a sufficient condition to preclude "roll up" of the invariant manifold at a sink; see the example in [2, p. 63]. The inequality $5M > \operatorname{Lip}(p)$ does not seem to have a geometric interpretation; rather it arises from the technical estimates in the proof.

We note that Robinson and Murdock in [10] prove the existence of invariant tori for a differential equation similar to system (2.9). Their result concerns the continuation of certain nonresonant unperturbed tori in analytic systems.

3. Normal hyperbolicity and continuation. In this section we recall the definition of normal hyperbolicity and discuss the basic idea, introduced by Kopell [8], that we will use to continue invariant manifolds. While our continuation method applies to normally hyperbolic invariant manifolds with expanding and contracting normal directions, in this paper we discuss only the case of normally hyperbolic invariant manifolds with no unstable normal directions. In particular, when we use the phrase "normally hyperbolic" we will use it in this restricted sense.

Let us consider a smooth differential equation

$$\dot{x} = F(x), \qquad x \in \mathbb{R}^n$$

with flow ϕ^t that has an overflowing invariant manifold $\overline{M} = M \cup \partial M$. Also, let TM denote the tangent bundle of M, and, with respect to the usual inner product on $\mathbb{R}^n$, let N denote the bundle normal to TM over M. Then,

$$T_M\mathbb{R}^n = TM \oplus N,$$

and there is a natural orthogonal projection $\Pi : T_M\mathbb{R}^n \to N$. Recall that in this context (see [3]) there are operators

$$\begin{aligned} &A_t(p) := D\phi^{-t}(p)|_{T_pM} : T_pM \to T_{\phi^{-t}(p)}M, \\ &B_t(p) := \Pi_p D\phi^t(\phi^{-t}(p))|_{N_{\phi^{-t}(p)}} : N_{\phi^{-t}(p)} \to N_p, \end{aligned} \tag{3.1}$$

and Lyapunov-type numbers, introduced in [3], are assigned to each point $p \in M$ as follows:

$$\nu(p) := \limsup_{t\to\infty} \|B_t(p)\|^{1/t}, \qquad \sigma(p) := \limsup_{t\to\infty} \frac{\ln \|A_t(p)\|}{-\ln \|B_t(p)\|}. \tag{3.2}$$

The number $\nu(p)$ measures the "exponential of the normal contraction rate" while $\sigma(p)$ compares the normal and tangential contraction rates. Both of these numbers are constant on orbits. Moreover, the Lyapunov-type numbers of an orbit are dominated by the supremum of the Lyapunov-type numbers on its α-limit set. Thus, to prove that M is normally hyperbolic, it suffices to compute the type numbers on the limit sets of the flow that are contained in M. A basic persistence result of Fenichel [3] states that if for all $p \in M$, we have $\nu(p) < 1$ and $\sigma(p) < 1/k$ for some positive integer k, then the manifold M persists under small C^1 perturbations by C^k vector fields. Moreover, the perturbed manifold is C^k. Let us mention that M, with the hypotheses of Fenichel's theorem is called k-normally hyperbolic. We will also use an equivalent formulation of k-normal hyperbolicity introduced by Hirsch, Pugh, and Shub [7]. A specialization of their definition to our perturbation problem is given below in display (8.11).

The persistence results just mentioned are widely applicable. However, in the perturbation problem (1.1) mentioned in the introduction, the existence of an invariant manifold cannot be obtained by a direct application of these persistence results due to the singular nature of the perturbation terms. Also, in the setting of the auxiliary family (1.3), the persistence result does not guarantee the existence of an invariant manifold up to $\delta = \epsilon$. On the other hand, in combination with an appropriate continuation method, the full strength of the persistence theory can be exploited to study perturbation problems of this type.

The idea of the continuation method is simple. To describe it, let us consider the smooth family E^ϵ of differential equations

$$\dot{x} = f(x, \epsilon). \tag{3.3}$$

Suppose that E^0 has a k-normally hyperbolic invariant manifold $M(0)$, and we wish to know if there is a corresponding family $M(\epsilon)$ of k-normally hyperbolic invariant manifolds that can be continued to some preassigned value of ϵ, say, $\epsilon = 1$. In this case, we can proceed in the following manner: Define A to be the set of all ϵ in the unit interval such that, for all $\epsilon' \in [0, \epsilon]$, the corresponding system $E^{\epsilon'}$ has a k-normally hyperbolic invariant manifold $M(\epsilon')$, and then prove that A is nonempty, open, and closed. Because $M(0)$ is a k-normally hyperbolic invariant manifold for E^0, A is not empty. The fact that A is open follows from the general persistence theory. Thus, all that remains is to show that A is closed; that is, if ϵ_* is the supremum of A, then $\epsilon_* \in A$. This can be accomplished in two steps: Prove that the system E^{ϵ_*} has a C^1 invariant manifold; then, prove that this invariant manifold is k-normally hyperbolic. Since we have a family of k-normally hyperbolic invariant manifolds $M(\epsilon)$ defined for $\epsilon' \in [0, \epsilon_*)$, the first step can be proved by showing that these manifolds are realized as graphs of an equicontinuous family of C^1 functions. The C^k smoothness of the limit manifold is obtained as a consequence of the second step which can be proved by checking the definition of k-normal hyperbolicity.

Let us consider a general family of the form

$$\dot{x} = f(x, \epsilon) + \epsilon g(x, \epsilon),$$

where, for some $\delta_0 > 0$, the system $\dot{x} = f(x, \delta)$ has a normally hyperbolic invariant manifold for $0 < \delta \leq \delta_0$. Kopell [8] studies a model equation that can be viewed as a special case of this family. To apply the general continuation method just described, she introduces an auxiliary family, which in our more general context would be

$$\dot{x} = f(x, \delta) + \epsilon g(x, \epsilon).$$

For this auxiliary system, if $\delta \in (0, \delta_0)$, then there is some $\epsilon(\delta) > 0$ such that, for $0 \leq \epsilon < \epsilon(\delta)$, the corresponding member of the auxiliary family has a normally hyperbolic invariant manifold $M(\delta, \epsilon)$. The idea is to fix some $\delta > 0$ sufficiently small so that continuation of normally hyperbolic invariant manifolds relative to the parameter ϵ can be carried out all the way to $\epsilon = \delta$. If this continuation is possible, then the member of the original family corresponding to $\epsilon = \delta$ has an invariant manifold.

In [8] (see also Wiggins [12, pp. 168–170]), a continuation theorem is stated for a family of the type described above, but of a more special form. However, the strategy of the proof of this theorem contains a gap. To describe the gap it is not necessary to consider the precise form of the equations or the hypotheses of the theorem. Rather, we will explain the problem in a general framework. Indeed, let us consider the parameter space of a family of differential equations and the subspace $\mathcal{N}$ corresponding to family members with a normally hyperbolic invariant manifold. Suppose that a path in $\mathcal{N}$ approaches the boundary of $\mathcal{N}$. Also, consider the supremum of each of the Lyapunov-type numbers ν and σ taken individually over the orbits of each normally hyperbolic invariant manifold in a continuous family. It is perhaps natural to suspect that the limit of at least one of these suprema converges to the number 1 as the path approaches the boundary. In other words, one might assume that the normal hyperbolicity is lost at the boundary only in this manner. However, this is not always the case. In fact, there may be paths for which the corresponding continuous family of normally hyperbolic invariant manifolds has both Lyapunov-type numbers uniformly bounded below one but the family of invariant manifolds does not converge to a C^1 manifold. Thus, in a continuation argument, it is required to prove that smooth invariant manifolds exist over the entire continuation interval and that all these manifolds are normally hyperbolic. The following example clearly shows why both requirements must be satisfied.

Consider a planar system

$$\dot{x} = f(x)$$

with a homoclinic loop at a hyperbolic saddle p whose eigenvalues α and β are such that $\alpha+\beta < 0$. In particular, the loop will be stable from the inside and the divergence of the vector field f at p is negative. Now add a one parameter family of perturbations $g(x,\epsilon)$ so that for $-1 < \epsilon < 0$ there is a limit cycle $\Gamma(\epsilon)$ that limits on the loop as ϵ approaches zero from the left and such that there is no limit cycle for $\epsilon > 0$. If we view $\Gamma(\epsilon)$ as an invariant manifold, then the corresponding Lyapunov-type number $\sigma(\epsilon)$ is identically zero. Also, the Lyapunov-type number $\nu(\epsilon)$ of $\Gamma(\epsilon)$ is exactly its Floquet multiplier; that is,

$$\nu(\epsilon) = e^{\frac{1}{T(\epsilon)}\int_0^{T(\epsilon)} \operatorname{div} f(\gamma(t,\epsilon))\, dt},$$

where $t \mapsto \gamma(t,\epsilon)$ is a periodic solution corresponding to the limit cycle and $T(\epsilon)$ is its period. Since the periodic solution spends most of its time near the hyperbolic saddle point, the Lyapunov-type number $\nu(\epsilon)$ approaches $e^{\alpha+\beta}$ as $\epsilon \to 0^-$. In particular, both Lyapunov-type numbers are bounded above by some number that is strictly less than one. But, the limit of the hyperbolic limit cycles is the nonsmooth homoclinic loop. Thus, in general, it is not enough to obtain uniform estimates on the Lyapunov-type numbers to ensure that a family of normally hyperbolic invariant manifolds can be continued.

4. Statement of main result. In this section we will state the continuation theorem that will be proved in this paper.

To obtain an invariant manifold for system (2.9) using a perturbation argument, it is useful to have an unperturbed system with an invariant manifold. As given, system (2.9), even after rescaling time, is degenerate in the limit as μ approaches zero. To remedy this problem, we will change coordinates and also rescale time so as to obtain a suitable perturbation problem.

Let us suppose that $\mu \neq 0$. Introduce new coordinates $\ell = \mu\hat{\rho}$, $\tau = \mu^2\varphi$, and a slow time $s = \mu^2 t$, and note that system (2.9) is equivalent to the system

$$\begin{aligned} \hat{\rho}' &= p(\sigma)\hat{\rho} + r(\sigma) + \mu^2 q(\sigma)\hat{\rho}^2 + \mu\widehat{R}(\mu\hat{\rho},\sigma,\tau/\mu^2,\mu),\\ \sigma' &= \lambda\hat{\rho} + g(\sigma) + \mu^2\nu\hat{\rho}^2 + \mu\widehat{S}(\mu\hat{\rho},\sigma,\tau/\mu^2,\mu),\\ \tau' &= 1, \end{aligned} \tag{4.1}$$

where the symbol “ $'$ ” denotes differentiation with respect to s. Let us also use the new coordinate $\rho := \lambda\hat{\rho} + g(\sigma)$ to express system (4.1) in the form

$$\begin{aligned} \rho' &= \Delta(\sigma)\rho + \Lambda(\sigma) + \mu R(\rho,\sigma,\tau/\mu^2,\mu),\\ \sigma' &= \rho + \mu S(\rho,\sigma,\tau/\mu^2,\mu),\\ \tau' &= 1, \end{aligned} \tag{4.2}$$

where

$$\Lambda(\sigma) := \lambda r(\sigma) - p(\sigma)g(\sigma), \qquad \Delta(\sigma) := p(\sigma) + g'(\sigma), \tag{4.3}$$

and the functions R and S are 2π periodic in σ and $2\pi m\mu^2/\Omega$ periodic in τ.

Let us write $\sigma \in \mathbb{S}^1$ to indicate that σ is an angular variable in the interval $0 \leq \sigma \leq 2\pi$ with the end points identified. Also, we will use the following hypothesis.

Hypothesis 1. *For each $\sigma \in \mathbb{S}^1$, $\Lambda(\sigma) \neq 0$ and $\Delta(\sigma) < 0$.*

Theorem 4.1. *If $k \geq 2$ is an integer, Hypothesis* 1 *holds, and $|\mu| > 0$ is sufficiently small, then system* (4.2) *has a k-normally hyperbolic invariant torus.*

We note that the k-normal hyperbolicity of the invariant torus in the conclusion of Theorem 4.1 implies that the invariant torus is C^k; see [7]. Also, as an immediate corollary of Theorem 4.1—just reverse the direction of time—the same conclusion holds under the assumption that, for each $\sigma \in \mathbb{S}^1$, $\Lambda(\sigma) \neq 0$ and $\Delta(\sigma) > 0$. Also, if we assume that $g(\sigma) \equiv 0$, as in Theorem 2.1 and if we assume that r has no zeros as in Hypothesis 1, then Theorem 4.1 is a generalization of Theorem 2.1. Indeed, the inequalities required in Theorem 2.1 are all replaced by the requirement that p has no zeros.

Finally, we mention that Theorem 4.1 is not valid if Hypothesis 1 is modified to allow the function Λ to have zeros. In fact, to obtain an analogue of Theorem 4.1 in case Λ has zeros, additional restrictions must be imposed. The formulation of the "right" hypotheses needed to prove an analogue of Theorem 4.1 in this case remains an interesting open problem.

The main idea of our proof of Theorem 4.1 is to view system (4.2) as a perturbation of the system

$$\begin{aligned} \rho' &= \Delta(\sigma)\rho + \Lambda(\sigma), \\ \sigma' &= \rho, \\ \tau' &= 1 \end{aligned} \tag{4.4}$$

and to show that the unperturbed system (4.4) has a normally hyperbolic invariant torus that continues to an invariant torus for system (4.2). We also note that the invariant torus for system (4.4) is the suspension of a normally hyperbolic invariant (simple closed) curve for the system

$$\begin{aligned} \rho' &= \Delta(\sigma)\rho + \Lambda(\sigma), \\ \sigma' &= \rho. \end{aligned} \tag{4.5}$$

5. Existence of an invariant curve. In this section we will prove that the unperturbed system (4.5) has a normally hyperbolic invariant curve. More precisely, we have the following theorem.

Theorem 5.1. *If Hypothesis* 1 *holds, then the system* (4.5) *has a C^∞ normally hyperbolic invariant simple closed curve given as the graph of a C^∞ function of the angular variable.*

There are several ways to prove Theorem 5.1. For example, a positively invariant annulus can be constructed, and the existence of a limit cycle can be proved using the Poincaré–Bendixson theorem. While the proof given below is more involved, it serves to illustrate the continuation technique that will be used in our proof of the existence of an invariant torus for system (4.1).

Our idea is to find a family of systems that includes system (4.5), to find a member of the family that has a normally hyperbolic invariant manifold, and then to continue this manifold through the family to the system (4.5).

Let us consider the family

$$\begin{aligned} \rho' &= \Delta(\sigma)\rho + \epsilon\Lambda(\sigma), \\ \sigma' &= \rho. \end{aligned} \tag{5.1}$$

We will use the next obvious lemma.

Lemma 5.2. *If Hypothesis* 1 *holds and* $\epsilon > 0$*, then system* (5.1) *has no rest point.*

Theorem 5.1 is an immediate consequence of the following lemma.

Lemma 5.3. *If Hypothesis* 1 *holds for system* (4.5)*, then system* (5.1) *has a* C^∞ *normally hyperbolic invariant curve that is given as the graph of a* C^∞ *function of the angular variable for all* $\epsilon \in [0, 1]$.

Proof. By Hypothesis 1, the function Λ does not vanish. Without loss of generality, we will assume that $\Lambda(\sigma) > 0$ for all σ. Also, by Hypothesis 1, the curve given by $\{(\rho, \sigma) : \rho = 0\}$ is a normally hyperbolic invariant manifold for the member of the family (5.1) at $\epsilon = 0$. Following the strategy discussed in section 3 let us consider the set A of all ϵ in the closed unit interval such that for all $\epsilon' \in [0, \epsilon]$ the corresponding member of the family (5.1) has a normally hyperbolic invariant closed curve γ^ϵ given as the graph of a C^∞ function h^ϵ of the angular variable. We will show that A is nonempty, open, and closed. This implies $A = [0, 1]$.

Because the invariant curve given by $\{(\rho, \sigma) : \rho = 0\}$ is normally hyperbolic for the family member at $\epsilon = 0$, we have that $0 \in A$ and therefore A is not empty. The fact that A is open follows from the persistence results for normally hyperbolic invariant manifolds. Let us define $\epsilon_* = \sup A$. To complete the proof we will show that $\epsilon_* \in A$; that is, A is closed.

Consider the family of curves

$$\Gamma(\kappa) := \{(\rho, \sigma) \in \mathbb{R}^2 : \rho - \kappa\Lambda(\sigma) = 0\},$$

where $\kappa \in \mathbb{R}$. Note that the curve $\Gamma(\kappa)$ is the graph of a periodic function of the angular variable. Thus, it separates the phase cylinder given by $(\rho, \sigma) \in \mathbb{R} \times \mathbb{S}^1$. Moreover, on $\Gamma(\kappa)$, by a straightforward computation, it follows that the dot product of the gradient of the function $(\rho, \sigma) \mapsto \rho - \kappa\Lambda(\sigma)$ and the vector field corresponding to the differential equation E^ϵ is given by

$$\left(-\kappa^2\Lambda'(\sigma) + \kappa\Delta(\sigma) + \epsilon\right)\Lambda(\sigma). \tag{5.2}$$

For $\epsilon \in [\epsilon_*/2, \epsilon_*)$, there exists $\kappa_0 > 0$ such that the coefficient of $\Lambda(\sigma)$ in (5.2) with $\kappa = \kappa_0$ is positive for all σ. Hence, the vector field corresponding to E^ϵ is transverse to the curve $\Gamma(\kappa_0)$. Because the function Λ is positive, it follows that γ^ϵ lies above the curve $\Gamma(\kappa_0)$; that is, $h^\epsilon(\sigma) > \kappa_0\Lambda(\sigma)$. Similarly, if $\nu \in \mathbb{R}$ is sufficiently large, then γ^ϵ lies below the curve $\{(\rho, \sigma) : \rho = \nu\}$. In particular, the set of functions $\mathcal{S} := \{h^\epsilon : \epsilon \in [\epsilon_*/2, \epsilon_*)\}$ is uniformly bounded.

Using the invariance, the function h^ϵ satisfies the differential equation

$$h^\epsilon_\sigma(\sigma) = \Delta(\sigma) + \epsilon\frac{\Lambda(\sigma)}{h^\epsilon(\sigma)}. \tag{5.3}$$

Thus, we have that $|h^\epsilon_\sigma| \leq |\Delta(\sigma)| + \epsilon_*/\kappa_0$ uniformly for $\epsilon \in [\epsilon_*/2, \epsilon_*)$, and, as a result, the set $\mathcal{S}$ is equicontinuous in the C^0 norm. By Arzela's theorem, there is a subsequence that converges to a continuous function h^{ϵ_*}.

We claim that the graph of h^{ϵ_*} is an invariant set for E^{ϵ_*}. To prove the claim, let $s \mapsto (\rho^\epsilon(s, q), \sigma^\epsilon(s, q))$ denote the solution of E^ϵ such that $\sigma^\epsilon(0, q) = q$ and $\rho^\epsilon(0, q) = h^\epsilon(q)$, and let us suppose that h^{ϵ_n} converges to h^{ϵ_*}. If $s \in \mathbb{R}$, then, using the continuity of the flow with respect to parameters, we have that $\sigma^{\epsilon_n}(s, q) \to \sigma^{\epsilon_*}(s, q)$ and $\rho^{\epsilon_n}(s, q) \to \rho^{\epsilon_*}(s, q)$. By passing to the limit as $n \to \infty$ in the identity

$\rho^{\epsilon_n}(s,q) = h^{\epsilon_n}(\sigma^{\epsilon_n}(s,q))$, we have $\rho^{\epsilon_*}(s,q) = h^{\epsilon_*}(\sigma^{\epsilon_*}(s,q))$. Thus, it follows that the graph of h^{ϵ_*} is an invariant set for E^{ϵ_*}. Because this invariant set is a single orbit of the differential equation, it is C^∞. Moreover, because the function Δ is everywhere negative, this invariant set is normally hyperbolic—it is a hyperbolic limit cycle. □

6. An a priori estimate for perturbed manifolds. The following proposition, which perhaps has independent interest, will play a key role in our proof of Theorem 4.1. While the statement of this proposition is natural, we do not know if it appears in the literature. Thus, we will give a complete proof in the appendix.

PROPOSITION 6.1. *Consider a smooth planar vector field*

$$x' = f(x) \tag{6.1}$$

with a periodic solution $t \mapsto x(t,p)$ of period ω corresponding to the periodic orbit Γ. If Γ is hyperbolic and asymptotically stable, that is,

$$b := \int_0^\omega \operatorname{tr} Df(x(t,p))\,dt < 0,$$

then there exist a neighborhood N of Γ and a constant $C > 0$ such that for every smooth function $g : N \to \mathbb{R}^2$ for which the differential equation

$$x' = f(x) + g(x) \tag{6.2}$$

has an invariant set $\bar{\Gamma} \subset N$, we have the following a priori estimate:

$$\sup\{d(x,\Gamma) : x \in \bar{\Gamma}\} \le C\|g\|_{C^0},$$

where $\|g\|_{C^0}$ is the supremum norm over N and d denotes the usual distance between sets.

PROPOSITION 6.2. *Consider a planar differential equation*

$$x' = f(x)$$

with a hyperbolic limit cycle Γ of period $T > 0$, and let τ be an angular variable modulo T. If Γ is asymptotically stable, then there is a neighborhood $N \subset \mathbb{R}^2 \times \mathbb{R}$ of the corresponding invariant torus M for the system

$$x' = f(x), \qquad \tau' = 1$$

and a constant $C > 0$ such that for every smooth function $g : N \to \mathbb{R}^2$, with $g(x, \tau + T) = g(x,\tau)$ for each $x \in \mathbb{R}^2$ and all $\tau \in \mathbb{R}$, and for which the system

$$x' = f(x) + g(x,\tau), \qquad \tau' = 1$$

has an invariant set $\bar{M} \subset N$, we have the a priori estimate

$$\sup\{d(x,M) : x \in \bar{M}\} \le C\|g\|_{C^0}.$$

7. Existence of invariant tori. In this section we will state our result on the existence of an invariant torus for the systems (2.9) and (4.2) as well as the main lemmas that we will use to prove it. In fact, we will prove the existence of invariant tori for systems of the more general form

$$\begin{aligned} \rho' &= f(\rho,\sigma) + \mu R(\rho,\sigma,\tau/\mu^2,\mu),\\ \sigma' &= g(\rho,\sigma) + \mu S(\rho,\sigma,\tau/\mu^2,\mu),\\ \tau' &= 1, \end{aligned} \tag{7.1}$$

where $|\mu| > 0$, where f, g, R, and S (redefined for this section) are all C^r functions that are 2π-periodic functions of the angular variable σ, and where the functions $t \mapsto R(\rho,\sigma,t,\mu)$ and $t \mapsto S(\rho,\sigma,t,\mu)$ are $2\pi m/\Omega$-periodic. For system (7.1), we view τ as an angular variable modulo $2\pi m\mu^2/\Omega$ and we let s denote the independent variable. Note that the slow time system (4.1) equivalent to system (2.9) is a special case of the differential (7.1). For a general discussion of integral manifolds for nonautonomous systems, see [1], [6].

To state our main result for system (7.1), let us consider the corresponding unperturbed system

$$\rho' = f(\rho,\sigma), \qquad \sigma' = g(\rho,\sigma), \tag{7.2}$$

and the following hypothesis.

HYPOTHESIS 2. *System* (7.2) *has an attracting hyperbolic limit cycle* Γ *that is the graph of a function of the angular variable* σ*.*

THEOREM 7.1. *If k is an integer such that $2 \le k \le r$ and the system* (7.2) *satisfies Hypothesis* 2*, then, for sufficiently small* $|\mu| > 0$*, system* (7.1) *has a k-normally hyperbolic invariant manifold that is the graph of a function of the angular variables σ and τ.*

The proof of Theorem 7.1 is given in the remaining sections of this paper using the lemmas that are stated below. Let us note at this point that it suffices to prove Theorem 7.1 for the case $\mu > 0$. The result for $\mu < 0$ follows from the first case by redefining the functions R and S in an obvious manner. Thus, we will consider only the case $\mu > 0$.

Let us consider the auxiliary family $E^{\epsilon,\mu}$ given by

$$\begin{aligned} \rho' &= f(\rho,\sigma) + \epsilon R(\rho,\sigma,\tau/\mu^2,\mu),\\ \sigma' &= g(\rho,\sigma) + \epsilon S(\rho,\sigma,\tau/\mu^2,\mu),\\ \tau' &= 1. \end{aligned} \tag{7.3}$$

Note that, by our assumption, the suspended system

$$\begin{aligned} \rho' &= f(\rho,\sigma),\\ \sigma' &= g(\rho,\sigma),\\ \tau' &= 1, \end{aligned} \tag{7.4}$$

where τ is viewed as a new angular variable modulo $2\pi m\mu^2/\Omega$, has a normally hyperbolic torus that is a graph over the two angular variables σ and τ. For our analysis we will consider the torus as a submanifold of the phase cylinder $\mathcal{C}$ given by $(\rho,\sigma,\tau) \in \mathbb{R}^3$, where σ and τ are viewed as the angular variables defined above. Topologically, $\mathcal{C}$ is the product of the real line with a two-dimensional torus.

For each $\mu > 0$, let us denote by A^μ the maximal interval with left endpoint at $\epsilon = 0$ such that the system $E^{\epsilon,\mu}$ has a k-normally hyperbolic invariant manifold, $k \geq 2$, as defined in [7]; see also display (8.11), given as the graph of a C^k function of the angular variables. Using the continuation strategy outlined in section 3, let us note that, for each $\mu > 0$, the set A^μ contains a nonempty relatively open interval with left endpoint $\epsilon = 0$. Moreover, if $\epsilon \in A^\mu$, then, by the general persistence results for normally hyperbolic invariant manifolds, there is an open interval containing ϵ that is contained in A^μ. Thus, Theorem 7.1 is an immediate consequence of the following proposition.

PROPOSITION 7.2. *Suppose that $\mu > 0$ and A^μ is the maximal interval with left endpoint at $\epsilon = 0$ such that the system $E^{\epsilon,\mu}$ has a k-normally hyperbolic invariant manifold, $k \geq 2$, that is the graph of a C^k function of the angular variables. If $\mu > 0$ is sufficiently small and if $\epsilon_* \leq \mu$ is the least upper bound of a relatively open interval with left endpoint $\epsilon = 0$ in A^μ, then $\epsilon_* \in A^\mu$.*

Proposition 7.2 is a consequence of the following three lemmas.

LEMMA 7.3. *With the hypotheses and notation of Proposition* 7.2*, the system $E^{\epsilon_*,\mu}$ has an invariant manifold $M(\epsilon_*, \mu)$ given as the graph of a C^1 function of the angular variables.*

LEMMA 7.4. *If $M(\epsilon_*, \mu)$ is the invariant manifold in Lemma* 7.3*, then it has an invariant normal bundle.*

LEMMA 7.5. *If $M(\epsilon_*, \mu)$ is the invariant manifold in Lemma* 7.3*, then $M(\epsilon_*, \mu)$ is k-normally hyperbolic. In particular, $M(\epsilon_*, \mu)$ is C^k and $\epsilon_* \in A^\mu$.*

8. Notation and preliminary results.

Lemmas (7.3)–(7.5) will be proved in the following sections. In this section we will define new notation and obtain some preliminary results that will be used in all three proofs. For the remainder of this section let us assume that $\mu > 0$ and $\epsilon \geq 0$ are fixed, and that system (7.3) has a normally hyperbolic invariant torus $M := M(\epsilon, \mu)$ given as the graph of the C^1 function h^ϵ of the angular variables.

8.1. Normal splitting and variational solutions.

The general results for normally hyperbolic invariant manifolds give the existence of an invariant splitting of $T_M\mathcal{C}$, the tangent bundle of the phase cylinder $\mathcal{C}$ restricted to this normally hyperbolic invariant torus, as a direct sum of the tangent bundle of the invariant torus M and an invariant normal bundle.

For notational convenience, let us define new functions

$$\begin{aligned} F(\rho, \sigma, \tau, \mu, \epsilon) &:= f(\rho, \sigma) + \epsilon R(\rho, \sigma, \tau/\mu^2, \mu), \\ G(\rho, \sigma, \tau, \mu, \epsilon) &:= g(\rho, \sigma) + \epsilon S(\rho, \sigma, \tau/\mu^2, \mu), \end{aligned} \tag{8.1}$$

and let us suppose that the invariant torus $M(\epsilon, \mu)$ is given as the graph of the function $(\sigma, \tau) \mapsto h^\epsilon(\sigma, \tau)$.

The vector field

$$\mathcal{X}_1^\epsilon(\sigma, \tau) := \begin{pmatrix} F(h^\epsilon(\sigma, \tau), \sigma, \tau, \mu, \epsilon) \\ G(h^\epsilon(\sigma, \tau), \sigma, \tau, \mu, \epsilon) \\ 1 \end{pmatrix} \tag{8.2}$$

is clearly tangent to $M(\epsilon, \mu)$. Also, as is easily seen by computing the tangents to each curve on $M(\epsilon,\mu)$ given by $\sigma \mapsto (h^\epsilon(\sigma,\tau),\sigma,\tau)$ for some fixed τ, the vector field

$$\mathcal{X}_2^\epsilon(\sigma,\tau) := \begin{pmatrix} h_\sigma^\epsilon(\sigma,\tau) \\ 1 \\ 0 \end{pmatrix} \tag{8.3}$$

is tangent to $M(\epsilon,\mu)$. Moreover, if $\xi = (h^\epsilon(\sigma,\tau),\sigma,\tau)$, then $\mathcal{X}_1^\epsilon(\sigma,\tau)$ and $\mathcal{X}_2^\epsilon(\sigma,\tau)$ span the corresponding fiber $T_\xi M(\epsilon,\mu)$ of the tangent bundle of $M(\epsilon,\mu)$.

To determine the contraction rates for the flow on the invariant torus $M(\epsilon,\mu)$, we must consider the solutions of the first variational equation for the system (7.3). If

$$s \mapsto \gamma^\epsilon(s,q) := (h^\epsilon(\sigma^\epsilon(s,q),\tau(s)),\sigma^\epsilon(s,q),\tau(s)) \tag{8.4}$$

is the solution of the system (7.3) with $\gamma^\epsilon(0,q) = (h^\epsilon(q,0),q,0)$, then the variational equation along the solution $s \mapsto \gamma^\epsilon(s,q)$ is given by

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} \begin{pmatrix} F_\rho & F_\sigma & F_\tau \\ G_\rho & G_\sigma & G_\tau \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \tag{8.5}$$

where the argument of each function in the system matrix is given by

$$(h^\epsilon(\sigma^\epsilon(s,q),\tau(s)),\sigma^\epsilon(s,q),\tau(s),\mu,\epsilon). \tag{8.6}$$

Proposition 8.1. *The variational* (8.5) *along the solution* (8.4) *on the invariant torus* $M(\epsilon,\mu)$ *has two independent solutions given by*

$$X_1(s) := \mathcal{X}_1^\epsilon(\gamma^\epsilon(s,q)), \qquad X_2(s) := y^\epsilon(s,q)\mathcal{X}_2^\epsilon(\gamma^\epsilon(s,q)), \tag{8.7}$$

where

$$y^\epsilon(t,q) := \exp\left(\int_0^t (G_\rho h_\sigma^\epsilon + G_\sigma)\,ds\right) \tag{8.8}$$

and the argument of F and G is given in display (8.6)*. Moreover, $X_1(s)$ and $X_2(s)$ span the tangent space of the invariant torus at each point along the solution* (8.4)*.*

Proof. The solution $X_1(s)$ is just the evaluation of the vector field corresponding to the base differential (7.3) along one of its integral curves. Thus, as is well known, it is a solution of the variational equation.

To obtain the second solution, let us recall that the invariant torus is given as a graph over the angular variables. In particular, the differential equation expressed in the corresponding local coordinates—the projection $(\rho,\sigma,\tau) \mapsto (\sigma,\tau)$ restricted to the graph is the coordinate map—is given by

$$\sigma' = G(h^\epsilon(\sigma,\tau),\sigma,\tau,\mu,\epsilon), \quad \tau' = 1.$$

The corresponding variational equation has the form

$$v' = (G_\rho h_\sigma^\epsilon + G_\sigma)v + (G_\rho h_\tau^\epsilon + G_\tau)w, \quad w' = 0.$$

One of its solutions is given by

$$s \mapsto (v(s),w(s)) = (y^\epsilon(s,q),0).$$

As $\rho = h^\epsilon(\sigma,\tau)$ on the invariant torus, the corresponding solution of the variational equation in the original coordinates is given by $y^\epsilon(s,q)\mathcal{X}_2^\epsilon(\gamma^\epsilon(s,q))$—substitute the general base solution into the local coordinate representation and then differentiate with respect to the initial condition.

In view of the fact that $\mathcal{X}_1^\epsilon$ and $\mathcal{X}_2^\epsilon$ are independent at each point of the manifold, and by virtue of the fact that y^ϵ is a positive function, the two solutions $X_1(s)$ and $X_2(s)$ are independent at each point along the solution γ^ϵ. □

Let $\Phi^\epsilon(s)$ denote the principal fundamental matrix solution of the variational equation (8.5) at $s = 0$. By the general theory of normally hyperbolic invariant manifolds, there is a normal bundle over the invariant torus M that is invariant under $\Phi^\epsilon(s)$. Because, M has codimension one, the fiber dimension of the normal bundle is one. Also, let us consider the family of cylinders given by

$$\mathcal{L}^s := \{(\rho,\sigma,\tau) : \tau = s\},$$

and note that $\mathcal{L} := \cup\{\mathcal{L}^s : s \in \mathbb{R}\}$ is a foliation of the phase cylinder that is invariant under the flow of system (7.3). Thus, it follows that $\mathcal{L}$ is also invariant for the variational equation, or equivalently, it is invariant under $\Phi^\epsilon(s)$. Because of the invariance of this foliation and the normal hyperbolicity, the fiber of the invariant normal bundle must be tangent to the leaf of this foliation that passes through the base point of the fiber. Also, the normal bundle of the embedded torus is trivial. Thus, it has a continuous nonzero section $\mathcal{X}_0^\epsilon$. Let us define

$$X_0(s) := \mathcal{X}_0^\epsilon(\gamma^\epsilon(s,q)), \tag{8.9}$$

where γ^ϵ is the solution defined in display (8.4). We remark here that the invariant normal bundle is required only to be continuous. In fact, in general it is not smooth.

To determine the growth rates required for the normal hyperbolicity, let us define

$$\lambda_1(s) := \frac{|X_1(s)|}{|X_1(0)|}, \quad \lambda_2(s) := \frac{|X_2(s)|}{|X_2(0)|}, \quad \lambda_3(s) := \frac{|\Phi^\epsilon(s)X_0(0)|}{|X_0(0)|}. \tag{8.10}$$

If k is a positive integer, then the invariant torus $M(\epsilon,\mu)$ is k-normally hyperbolic, as defined in [7], provided that there are numbers $\beta > 0$ and $c > 0$ independent of the choice of the solution on $M(\epsilon,\mu)$ such that the following conditions are satisfied for $s \geq 0$:

$$\lambda_3(s) \leq ce^{-\beta s}, \quad \frac{\lambda_3(s)}{\lambda_1^k(s)} \leq ce^{-\beta s}, \quad \frac{\lambda_3(s)}{\lambda_2^k(s)} \leq ce^{-\beta s}. \tag{8.11}$$

8.2. A formula for $\boldsymbol{\lambda_3(s)}$. The vector function $s \mapsto X_2(s)$ defined in display (8.7) is a solution of the system (8.5). Define

$$X_2^\perp(s) := \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} X_2(s),$$

and note that there are smooth functions $s \mapsto a(s)$ and $s \mapsto b(s)$ such that

$$\Phi^\epsilon(s)X_2^\perp(0) = a(s)X_2(s) + b(s)X_2^\perp(s).$$

Moreover, it is not difficult to compute the following formulas:

$$\begin{aligned} b(s) &= \frac{|X_2(0)|^2}{|X_2(s)|^2} \exp\left(\int_0^s \operatorname{tr} B(t)\,dt\right),\\ a'(s) &= \frac{b(s)}{|X_2(s)|^2}(\langle B(s)X_2(s), X_2^{\perp}(s)\rangle + \langle B(s)X_2^{\perp}(s), X_2(s)\rangle),\\ a(0) &= 0, \end{aligned} \tag{8.12}$$

where $B(s)$ is the system matrix of the linear system (8.5).

By the above remarks, the vector $X_0(s)$ is in the span of the linearly independent vectors $X_2(s)$ and $X_2^{\perp}(s)$. Thus, by an appropriate choice of the nonzero normal bundle section $\mathcal{X}_0^{\epsilon}$, there is a smooth function $s \mapsto \alpha(s)$ such that

$$X_0(s) = \alpha(s)X_2(s) + X_2^{\perp}(s) \tag{8.13}$$

and a smooth function $s \mapsto \lambda(s)$ such that

$$\Phi^{\epsilon}(s)X_0(0) = \lambda(s)X_0(s). \tag{8.14}$$

By substitution of the identity (8.13) into (8.14) and by using the independence of X_2 and $X_2^{\perp}$, it follows that $\lambda(s)\alpha(s) = \alpha(0) + a(s)$ and $\lambda(s) = b(s)$. Hence, using the definition given in display (8.10), and the formulas obtained in this section, we have the following equalities:

$$\begin{aligned} \lambda_3(s) = \lambda(s)\frac{|X_0(s)|}{|X_0(0)|} &= \left(\frac{|X_2(0)|^2}{|X_2(s)|^2} \exp\left(\int_0^s \operatorname{tr} B(t)\,dt\right)\right) \frac{(\alpha^2(s)+1)^{1/2}|X_2(s)|}{(\alpha^2(0)+1)^{1/2}|X_2(0)|}\\ &= \frac{|X_2(0)|}{|X_2(s)|}\left(\frac{\alpha^2(s)+1}{\alpha^2(0)+1}\right)^{1/2} \exp\left(\int_0^s \operatorname{tr} B(t)\,dt\right). \end{aligned} \tag{8.15}$$

8.3. Derivative estimates. Under the assumptions that $\mu > 0$ and the unperturbed system (7.2) have a normally hyperbolic invariant manifold given as a graph of a function of the angular variables, we know that $E^{\epsilon,\mu}$, for sufficiently small $\epsilon > 0$, has a normally hyperbolic invariant manifold given as the graph of a function h^{ϵ} of the angular variables. In this section, we will determine some a priori estimates on the size of the derivatives of h^{ϵ}. We will first state and prove two lemmas. The first reduces the main estimate from the vector to the scalar case, while the second lemma gives certain properties of an operator equation for one of the derivatives that must be estimated.

8.3.1. A reduction lemma. The next lemma shows that it suffices to estimate the partial derivative h_{σ} on a Poincaré section.

LEMMA 8.2. *Suppose that $\mu > 0$ and $\epsilon_* > 0$ and that $E^{\epsilon,\mu}$ has an invariant manifold given as the graph of the function h^{ϵ} of the angular variables for $0 \le \epsilon < \epsilon_* \le \mu$. If there is a constant $C_1 > 0$ such that, for all angles σ and τ, the following estimates hold:*

$$|h^{\epsilon}(\sigma,\tau) - h^0(\sigma,\tau)| < C_1\epsilon, \qquad |h^{\epsilon}_{\sigma}(\sigma,0) - h^0_{\sigma}(\sigma,0)| < C_1\epsilon,$$

then for μ sufficiently small there is a constant $C_2 > 0$ such that

$$|h^{\epsilon}_{\sigma}(\sigma,\tau) - h^0_{\sigma}(\sigma,\tau)| < C_2\epsilon.$$

Proof. Let us suppose first that $\dot{x} = F(x, \epsilon)$ is a smooth family of differential equations. If x^ϵ and x^0 are solutions of the members of this family corresponding to their superscripts, then by an application of Gronwall's inequality there is a constant $K > 0$ such that

$$|x^\epsilon(t) - x^0(t)| < K e^{K|t|}(|x^\epsilon(0) - x^0(0)| + \epsilon|t|). \tag{8.16}$$

Let $s \mapsto \phi(s, (\rho, \sigma, \tau), \epsilon)$ be the solution of the system $E^{\epsilon,\mu}$ with the initial condition $\phi(0, (\rho, \sigma, \tau), \epsilon) = (\rho, \sigma, \tau)$, and note that the solution γ^ϵ defined in display (8.4) is given by $\gamma^\epsilon(s, q) = \phi(s, (h^\epsilon(q, 0), q, 0), \epsilon)$. For each pair of angles p and τ with $0 \le \tau < 2\pi m\mu^2/\Omega$, there is a unique angle q^ϵ defined by the equation

$$(h^\epsilon(q^\epsilon, 0), q^\epsilon, 0) = \phi(-\tau, (h^\epsilon(p, \tau), p, \tau), \epsilon). \tag{8.17}$$

By an application of the inequality (8.16) to the family of solutions (8.17), if μ is sufficiently small, then there is a constant $K_1 > 0$, that does not depend on the choice of τ, such that

$$\begin{aligned} |h^\epsilon(q^\epsilon, 0) - h^0(q^0, 0)| + |q^\epsilon - q^0| &\le K_1(|h^\epsilon(p, \tau) - h^0(p, \tau)| + \epsilon) \\ &\le K_1(C_1 + 1)\epsilon. \end{aligned} \tag{8.18}$$

In particular, there is a constant $K_2 > 0$ such that

$$|q^\epsilon - q^0| \le K_2\epsilon. \tag{8.19}$$

By inverting the flow in (8.17), we have that $\gamma^\epsilon(\tau, q^\epsilon) = (h^\epsilon(p, \tau), p, \tau)$. Using an obvious modification of the notation as well as the result of Proposition 8.1, let us consider the first variational equation for $E^{\epsilon,\mu}$ along the solution $s \mapsto \gamma^\epsilon(s, q^\epsilon)$ and the solution of this variational equation that is given by

$$X_2^\epsilon(s, q^\epsilon) = y^\epsilon(s, q^\epsilon) \begin{pmatrix} h_\sigma^\epsilon(\sigma^\epsilon(s, q^\epsilon), \tau(s)) \\ 1 \\ 0 \end{pmatrix}.$$

Note that its initial condition and its value at $s = \tau$ are given by

$$X_2^\epsilon(0, q^\epsilon) = \begin{pmatrix} h_\sigma^\epsilon(q^\epsilon, 0) \\ 1 \\ 0 \end{pmatrix}, \qquad X_2^\epsilon(\tau, q^\epsilon) = y^\epsilon(\tau, q^\epsilon) \begin{pmatrix} h_\sigma^\epsilon(p, \tau) \\ 1 \\ 0 \end{pmatrix}.$$

By an application of the inequality (8.16) to this family of solutions of the variational equation, if μ is sufficiently small, then there is a constant $K_2 > 0$ such that

$$\begin{aligned} |y^\epsilon(\tau, q^\epsilon)h_\sigma^\epsilon(p, \tau) &- y^0(\tau, q^0)h_\sigma^0(p, \tau)| + |y^\epsilon(\tau, q^\epsilon) - y^0(\tau, q^0)| \\ &\le K_2(|h_\sigma^\epsilon(q^\epsilon, 0) - h_\sigma^0(q^0, 0)| + \epsilon) \\ &\le K_2(|h_\sigma^\epsilon(q^\epsilon, 0) - h_\sigma^0(q^\epsilon, 0)| + |h_\sigma^0(q^\epsilon, 0) - h_\sigma^0(q^0, 0)| + \epsilon). \end{aligned}$$

Moreover, by the hypothesis of the lemma, by inequality (8.19), and by the fact that h_σ^0 is Lipschitz, we have that there is a constant $K_3 > 0$ such that

$$|y^\epsilon(\tau, q^\epsilon)h_\sigma^\epsilon(p, \tau) - y^0(\tau, q^0)h_\sigma^0(p, \tau)| + |y^\epsilon(\tau, q^\epsilon) - y^0(\tau, q^0)| \le K_3\epsilon. \tag{8.20}$$

In particular, both summands on the left-hand side of the last inequality are bounded above by $K_3\epsilon$.

By a reverse triangle law estimate starting with the fact that the quantity

$$|(y^\epsilon(\tau,q^\epsilon)h^\epsilon_\sigma(p,\tau)-y^\epsilon(\tau,q^\epsilon)h^0_\sigma(p,\tau))-(y^0(\tau,q^0)h^0_\sigma(p,\tau)-y^\epsilon(\tau,q^\epsilon)h^0_\sigma(p,\tau))|$$

is bounded above by $K_3\epsilon$, by the inequality (8.20), and the fact that h^0_σ is uniformly bounded, we find that there is a constant $K_4>0$ such that

$$|y^\epsilon(\tau,q^\epsilon)||h^\epsilon_\sigma(p,\tau)-h^0_\sigma(p,\tau)|\le K_3\epsilon+|h^0_\sigma(p,\tau)||y^\epsilon(\tau,q^\epsilon)-y^0(\tau,q^0)|\le K_4\epsilon. \tag{8.21}$$

By a second reverse triangle law estimate, we have that

$$|y^\epsilon(\tau,q^\epsilon)|=|y^0(\tau,q^0)+(y^\epsilon(\tau,q^\epsilon)-y^0(\tau,q^0))|\ge|y^0(\tau,q^0)|-K_3\mu.$$

Also, if we take $\mu>0$ sufficiently small, then there is a constant $K_5>0$ such that $|y^\epsilon(\tau,q^\epsilon)|>K_5$. The result follows from this fact and the inequality (8.21). □

8.3.2. The estimates. Some of the most important estimates that are required for the proof of our main result are given in the next lemma.

LEMMA 8.3. *Suppose Hypothesis* 2 *holds for the unperturbed system* (7.2). *There is a number* $\mu>0$ *and a constant* $C>0$ *such that if* ϵ_* *is as in Proposition* 7.2, *then*

$$|h^\epsilon-h^0|_{C^0}<C\epsilon,\quad |h^\epsilon_\sigma-h^0_\sigma|_{C^0}\le C\epsilon,\quad |h^\epsilon_\tau|_{C^0}\le C\epsilon$$

for $0\le\epsilon<\epsilon_*$.

Proof. We will show that, if $\mu>0$ is sufficiently small, then there is a constant $C>0$ such that $|h^\epsilon_\sigma-h^0_\sigma|_{C^0}\le C\epsilon$. By Lemma 8.2, it suffices to find a constant $C>0$ such that, for all $q\in\mathbb{S}^1$,

$$|h^\epsilon_\sigma(q,0)-h^0_\sigma(q,0)|\le C\epsilon.$$

To prove this inequality, let us recall Proposition 8.1, and note that the function given by $s\mapsto(h^\epsilon_\sigma(\sigma^\epsilon(s,q),\tau(s))y^\epsilon(s,q),y^\epsilon(s,q))$ is a solution of the "subsystem" of system (8.5) given by

$$\begin{aligned}u'&=(f_\rho+\epsilon R_\rho)u+(f_\sigma+\epsilon R_\sigma)v,\\ v'&=(g_\rho+\epsilon S_\rho)u+(g_\sigma+\epsilon S_\sigma)v.\end{aligned}\tag{8.22}$$

If $\Psi^\epsilon(s,q):=(\psi^\epsilon_{ij}(s,q))_{2\times2}$ is the principal fundamental matrix solution of (8.22) at $s=0$, then

$$\begin{pmatrix}h^\epsilon_\sigma(\sigma^\epsilon(s,q),\tau(s))y^\epsilon(s,q)\\ y^\epsilon(s,q)\end{pmatrix}=\Psi^\epsilon(s,q)\begin{pmatrix}h^\epsilon_\sigma(q,0)\\ 1\end{pmatrix}.$$

Therefore, we have

$$h^\epsilon_\sigma(\sigma^\epsilon(s,q),\tau(s))=\frac{\psi^\epsilon_{11}(s,q)h^\epsilon_\sigma(q,0)+\psi^\epsilon_{12}(s,q)}{\psi^\epsilon_{21}(s,q)h^\epsilon_\sigma(q,0)+\psi^\epsilon_{22}(s,q)}.\tag{8.23}$$

Because the second angular argument τ will often be set to $\tau=0$, in the remainder of the proof we will suppress the second argument in expressions involving the functions h^ϵ and their partial derivatives whenever $\epsilon>0$ and the second angular argument is set to $\tau=0$. In addition, the function h^0 is constant with respect to τ;

thus we will always suppress its second angular argument and write h^0 as a function of the first angular variable only.

If $s \mapsto (\rho(s), \sigma(s))$ is a solution of the unperturbed system

$$\rho' = f(\rho, \sigma), \qquad \sigma' = g(\rho, \sigma),$$

then the vector function

$$s \mapsto (f(h^0(\sigma), \sigma), g(h^0(\sigma), \sigma))$$

is a solution of the corresponding variational equation (8.22) with $\epsilon = 0$. Using the fact that $\rho(s) = h^0(\sigma(s))$ on the invariant manifold, and differentiating with respect to s, we find that

$$f(h^0(\sigma(s)), \sigma(s)) = h^0_\sigma(\sigma(s))g(h^0(\sigma(s)), \sigma(s)).$$

Thus, the function $s \mapsto (h^0_\sigma(\sigma(s))g(h^0(\sigma(s)), \sigma(s)), g(h^0(\sigma(s)), \sigma(s)))$ is a solution of the variational equation. Using the fundamental matrix solution Ψ^ϵ defined after display (8.22), we find that

$$\psi^0_{21}(s,q)h^0_\sigma(q) + \psi^0_{22}(s,q) = \frac{g(h^0(\sigma(s,q)), \sigma(s,q))}{g(h^0(q), \sigma(q))}.$$

In view of the hypothesis that σ' does not vanish, there is a number $M_0 > 0$ such that $|\psi^0_{21}(s,q)h^0_\sigma(q) + \psi^0_{22}(s,q)| \geq M_0$ for every $s \in \mathbb{R}$ and $q \in \mathbb{S}^1$.

By hypothesis, the unperturbed normally hyperbolic invariant manifold for $E^{0,\mu}$ is the suspension of an attracting hyperbolic limit cycle. In particular, the characteristic multiplier of the limit cycle is negative. Using the fact that the determinant of the fundamental matrix solution of the variational equation is proportional to the exponential of the integral of the divergence of the vector field evaluated along the limit cycle, it follows that there is some $T_0 > 0$ such that, for $t \geq T_0$, we have the inequality $\det \Psi^0(t,q) \leq (1/4)M_0^2$. If, in addition, $0 < \mu^2 < \Omega/(2\pi m)$, then there is a positive integer n such that $T_0 \leq 2\pi m n\mu^2/\Omega < T_0 + 1$. For definiteness, let $n = n(\mu)$ denote the smallest such integer, and define

$$T := T(\mu) = 2\pi m n \mu^2/\Omega. \tag{8.24}$$

While T will vary as $\mu > 0$ is made sufficiently small so that new requirements are satisfied, the final value of T is an integer multiple of the period of the perturbation terms in the corresponding differential equation $E^{\epsilon,\mu}$, the value of T is bounded above and below, and T approaches T_0 as μ decreases toward zero.

If we set $s = T$ in (8.23), then

$$h^\epsilon_\sigma(p) = \frac{\psi^\epsilon_{11}(T,q^\epsilon)h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{12}(T,q^\epsilon)}{\psi^\epsilon_{21}(T,q^\epsilon)h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{22}(T,q^\epsilon)}, \tag{8.25}$$

where q^ϵ is defined by the relation

$$p = \sigma^\epsilon(T, q^\epsilon). \tag{8.26}$$

Choose a bounded neighborhood N of the graph of h^0 and the corresponding constant $C_0 > 0$ as in Proposition 6.2. Also, choose $r_0 > 0$ so small that if $|h^\epsilon - h^0|_{C^0} < r_0$, then the graph of h^ϵ is in N, and note that

$$K := \sup\left\{|R(\rho, \sigma, \tau/\mu^2, \mu)| + |S(\rho, \sigma, \tau/\mu^2, \mu)| : (\rho, \sigma, \tau) \in N,\ 0 < \mu^2 < \frac{\Omega}{2\pi m}\right\}$$

is finite—the functions R and S are the perturbation terms of the system $E^{\epsilon,\mu}$ in display (7.3).

Choose $r > 0$ sufficiently small so that, for $T_0 \leq T < T_0 + 1$,

$$|d_{21}\xi + d_{22}| > \frac{2}{3}M_0, \qquad |\det D - \det \Psi^0(T, q)| < \frac{1}{8}M_0^2 \tag{8.27}$$

whenever ξ, a real number, D, a 2×2 real matrix, and $q \in \mathbb{S}^1$ are such that

$$|\xi - h^0_\sigma(q)| < r, \qquad |D - \Psi^0(T, q)| < r.$$

The existence of r with the required properties follows from the continuity of the map $(u, v, w) \mapsto |uv + w|$, the continuity of the determinant function, and the compactness of $\mathbb{S}^1$.

If we choose $\mu > 0$ so small that $0 < \mu^2 < \Omega/2\pi m$ and $C_0 K\mu < r_0$, then, by Proposition 6.2,

$$|h^\epsilon - h^0|_{C^0} < C_0 K\epsilon \tag{8.28}$$

as long as $|h^\epsilon - h^0|_{C^0} < r_0$. Thus, we have that the inequality (8.28) holds for $0 \leq \epsilon < \epsilon_*$. Also, by an application of the Gronwall estimate (8.16) applied to the solutions $\gamma^\epsilon(-s, p)$ and $\gamma^0(-s, p)$ defined in display (8.4), we find that there is a constant $C_1 > 0$ such that

$$|h^\epsilon(q^\epsilon) - h^0(q^0)| + |q^\epsilon - q^0| \leq C_1(|h^\epsilon(p) - h^0(p)| + \epsilon).$$

In view of the estimate in display (8.28), we conclude that there is a constant $C_2 > 0$ such that $|q^\epsilon - q^0| \leq C_2\epsilon$. By an application of the estimate (8.16) to the solutions $s \mapsto \Psi^\epsilon(s, q^\epsilon)$ and $s \mapsto \Psi^0(s, q^0)$ of the variational equation, we have, for some $C_3 > 0$, the inequality

$$|\Psi^\epsilon(T, q^\epsilon) - \Psi^0(T, q^0)| \leq C_3(|q^\epsilon - q| + \epsilon).$$

To obtain the estimates in the statement of the lemma, we will first prove the following claim.

Claim. There exists a constant $C > 0$ such that, for $0 \leq \epsilon < \epsilon_*$, if $|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \leq r$, then $|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \leq C\epsilon$.

Proof of claim. Fix $p \in \mathbb{S}^1$, and let q^ϵ be as in (8.26). Using this notation and the identity (8.25), we have

$$\begin{aligned}|h^\epsilon_\sigma(p) - h^0_\sigma(p)| \leq{}& \left|\frac{\psi^\epsilon_{11}(T, q^\epsilon)h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{12}(T, q^\epsilon)}{\psi^\epsilon_{21}(T, q^\epsilon)h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{22}(T, q^\epsilon)} - \frac{\psi^\epsilon_{11}(T, q^\epsilon)h^0_\sigma(q^\epsilon) + \psi^\epsilon_{12}(T, q^\epsilon)}{\psi^\epsilon_{21}(T, q^\epsilon)h^0_\sigma(q^\epsilon) + \psi^\epsilon_{22}(T, q^\epsilon)}\right| \\ &+ \left|\frac{\psi^\epsilon_{11}(T, q^\epsilon)h^0_\sigma(q^\epsilon) + \psi^\epsilon_{12}(T, q^\epsilon)}{\psi^\epsilon_{21}(T, q^\epsilon)h^0_\sigma(q^\epsilon) + \psi^\epsilon_{22}(T, q^\epsilon)} - \frac{\psi^0_{11}(T, q^0)h^0_\sigma(q^0) + \psi^0_{12}(T, q^0)}{\psi^0_{21}(T, q^0)h^0_\sigma(q^0) + \psi^0_{22}(T, q^0)}\right| \\ :={}& I + II.\end{aligned}$$

To estimate the quantities I and II, let us consider, for real numbers ξ and 2×2 matrices $D = (d_{ij})$, the function $u : \mathbb{R} \times \mathbb{R}^4 \to \mathbb{R}$ defined by

$$u(\xi, D) = \frac{d_{11}\xi + d_{12}}{d_{21}\xi + d_{22}}.$$

Using this function, we have

$$\begin{aligned} I &= |u(y, \Psi^\epsilon(T, q^\epsilon)) - u(x, \Psi^\epsilon(T, q^\epsilon))|, \\ II &= |u(x, \Psi^\epsilon(T, q^\epsilon)) - u(z, \Psi^0(T, q^0))|, \end{aligned}$$

where $x := h^0_\sigma(q^\epsilon)$, $y := h^\epsilon_\sigma(q^\epsilon)$, and $z := h^0_\sigma(q)$.

To estimate I, apply the mean value theorem to the function $\xi \mapsto u(\xi, D)$ and use the fact that

$$u_\xi(\xi, D) = \frac{\det D}{d_{21}\xi + d_{22}}$$

to obtain the inequality

$$I \le \left|\frac{\det \Psi^\epsilon(T, q^\epsilon)}{\psi^\epsilon_{21}(T, q^\epsilon)\xi + \psi^\epsilon_{22}(T, q^\epsilon)}\right| |h^\epsilon_\sigma(q^\epsilon) - h^0_\sigma(q^\epsilon)|$$

for some ξ between $h^\epsilon_\sigma(q^\epsilon)$ and $h^0_\sigma(q^\epsilon)$. Let us note that $|\xi - h^0_\sigma(q^\epsilon)| < r$. Also, if $\mu > 0$ is sufficiently small, then $|\Psi^\epsilon(T, q^\epsilon) - \Psi^0(T, q^0)|_{C^0} < r$. Thus, if we use the inequalities (8.27) together with a triangle law estimate for the term containing the determinant, then we find that $I \le \lambda |h^\epsilon_\sigma - h^0_\sigma|_{C^0}$ for $\lambda = 27/32 < 1$.

To estimate II, let us note that the function u is Lipschitz on the set

$$\{(\xi, D) : \xi = h^0_\sigma(q),\ D = \Psi^\epsilon(T, q),\ q \in \mathbb{S}^1,\ 0 \le \epsilon < \epsilon^*\}.$$

In particular, there is a constant $L_1 > 0$ such that

$$II \le L_1(|h^0_\sigma(q^\epsilon) - h^0_\sigma(q^0)| + |\Psi^\epsilon(T, q^\epsilon) - \Psi^0(T, q^0)|).$$

Using the fact that h^0 is Lipschitz on $\mathbb{S}^1$, we conclude that there exist constants $L > 0$ and $C_4 > 0$ such that

$$II \le L(|q^\epsilon - q^0| + |\Psi^\epsilon(T, q^\epsilon) - \Psi^0(T, q^0)|) \le C_4\epsilon.$$

Thus,

$$|h^\epsilon_\sigma(p) - h^0_\sigma(p)| \le C_4\epsilon + \lambda|h^\epsilon_\sigma(q) - h^0_\sigma(q)| \le C_4\epsilon + \lambda|h^\epsilon_\sigma - h^0_\sigma|_{C^0},$$

and, as a result,

$$|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \le \frac{C_4}{1-\lambda}\epsilon.$$

This completes the proof of the claim.

In addition to the restrictions on the size of μ already made, let us also require that $\mu < r/C$, where C is the constant appearing in the claim. Define

$$\epsilon_0 = \sup\{\epsilon' : 0 \le \epsilon' < \epsilon_*,\ |h^\epsilon_\sigma - h^0_\sigma|_{C^0} \le r \text{ for } \epsilon \in [0, \epsilon']\}.$$

We will show that $\epsilon_0 = \epsilon_*$. Suppose not, then $\epsilon_0 < \epsilon_*$. For $\epsilon < \epsilon_0$, $|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \le r$ and, hence, $|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \le C\epsilon$ by the claim. Since $\epsilon_0 < \epsilon_*$, the graph of h^{ϵ_0} is normally hyperbolic by the definition of ϵ_*. Passing to the limit as $\epsilon \to \epsilon_0$, we have $|h^{\epsilon_0}_\sigma - h^0_\sigma|_{C^0} \le C\epsilon_0 < r$. This contradicts the fact that ϵ_0 is the supremum; hence $\epsilon_0 = \epsilon_*$. Now, by the claim, we conclude that $|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \le C\epsilon$ for $0 \le \epsilon < \epsilon_*$.

Let us now estimate h^ϵ_τ. Note first that the invariance of the graph of the function h^ϵ is equivalent to the identities

$$\begin{aligned} h^\epsilon_\sigma \sigma' + h^\epsilon_\tau &= f(h^\epsilon, \sigma) + \epsilon R(h^\epsilon, \sigma, \tau/\mu^2, \mu), \\ \sigma' &= g(h^\epsilon, \sigma) + \epsilon S(h^\epsilon, \sigma, \tau/\mu^2, \mu). \end{aligned} \tag{8.29}$$

If we suppose that $h^\epsilon(\sigma, \tau) = h^0(\sigma) + \epsilon H(\sigma, \tau, \epsilon)$ and substitute this expression into the relation (8.29), then we obtain the equation

$$\begin{aligned} &(g(h^0 + \epsilon H, \sigma) + \epsilon S(h^0 + \epsilon H, \sigma, \tau/\mu^2, \mu))H_\sigma + H_\tau \\ &\qquad = \frac{1}{\epsilon}(f(h^0 + \epsilon H, \sigma) - h^0_\sigma g(h^0 + \epsilon H, \sigma)) - h^0_\sigma S + R. \end{aligned}$$

Finally, using the estimate

$$|H_\sigma|_{C^0} = \frac{1}{\epsilon}|h^\epsilon_\sigma - h^0_\sigma|_{C^0} \le C_1$$

and the relation $f(h^0, \sigma) = h^0_\sigma g(h^0, \sigma)$, we conclude that there is a constant $C > 0$, that is independent of ϵ, such that $|H_\tau|_{C^0} \le C$; that is, $|h^\epsilon_\tau|_{C^0} \le C\epsilon$. □

Let us note that we will eventually have to verify estimates as in display (8.11). As a step in this direction, let us first consider λ_3 and note, from the formula (8.15), that the growth estimate requires an asymptotic estimate of the norm of $X_2(s)$. By the definition of $X_2(s)$, it is clear that this norm estimate is determined by the behavior of the function $s \mapsto y^\epsilon(s, q)$ defined in display (8.8). To determine this behavior, we must estimate the integral

$$\int_0^s (G_\rho h^\epsilon_\sigma + G_\sigma)\, dt. \tag{8.30}$$

The precise integral estimate that we will require is the content of the next lemma.

Lemma 8.4. *If, for all $s \ge 0$, the function $s \mapsto G(h^\epsilon(\sigma(s), \tau(s)), \sigma(s), \tau(s), \epsilon)$ has no zeros, and if $\mu > 0$ is sufficiently small, then there is a constant $C > 0$ such that*

$$y^\epsilon(s, q) = \exp\left(\int_0^s (G_\rho h^\epsilon_\sigma + G_\sigma)\, dt\right) \ge e^{-C} e^{-C\epsilon s}.$$

Proof. Note that

$$\begin{aligned} \frac{d}{ds} \ln |G(h^\epsilon(\sigma(s), \tau(s)), \sigma(s), \tau(s), \epsilon)| &= \frac{1}{G}(G_\rho h^\epsilon_\sigma \sigma' + G_\rho h^\epsilon_\tau + G_\sigma \sigma' + G_\tau) \\ &= G_\rho h^\epsilon_\sigma + G_\sigma + \frac{G_\rho h^\epsilon_\tau}{G} + \frac{G_\tau}{G}. \end{aligned}$$

After integration over the interval $[0, s]$ and a rearrangement, we obtain the identity

$$\begin{aligned} \int_0^s (G_\rho h^\epsilon_\sigma + G_\sigma)\, dt = \ln \Big| &\frac{G(h^\epsilon(\sigma(s), \tau(s)), \sigma(s), \tau(s), \epsilon)}{G(h^\epsilon(\sigma(0), \tau(0)), \sigma(0), \tau(0), \epsilon)} \Big| \\ &- \int_0^s \frac{G_\rho h^\epsilon_\tau}{G}\, dt - \int_0^s \frac{G_\tau}{G}\, dt. \end{aligned} \tag{8.31}$$

Recall the definition (8.1) of G, and note that

$$\frac{d}{ds}\Big(\frac{S}{G}\Big) = \frac{S_\rho h^\epsilon_\sigma \sigma' + S_\rho h^\epsilon_\tau + S_\sigma \sigma' + \frac{1}{\mu^2} S_\tau}{G} \\ - \frac{S(G_\rho h^\epsilon_\sigma \sigma' + G_\rho h^\epsilon_\tau + G_\sigma \sigma' + \frac{\epsilon}{\mu^2} S_\tau)}{G^2}.$$

Using this identity and an easy computation, we find that

$$\begin{aligned}\int_0^s \frac{G_\tau}{G}\,dt &= \frac{\epsilon}{\mu^2}\int_0^s \frac{S_\tau}{G}\,dt \\ &= \epsilon\Big(\frac{S(h^\epsilon(\sigma(s),\tau(s)),\sigma(s),\tau(s),\epsilon)}{G(h^\epsilon(\sigma(s),\tau(s)),\sigma(s),\tau(s),\epsilon)} - \frac{S(h^\epsilon(\sigma(0),\tau(0)),\sigma(0),\tau(0),\epsilon)}{G(h^\epsilon(\sigma(0),\tau(0)),\sigma(0),\tau(0),\epsilon)}\Big) \\ &\quad - \epsilon\int_0^s \frac{S_\rho h^\epsilon_\sigma G + S_\rho h^\epsilon_\tau + S_\sigma G}{G}\,dt \\ &\quad + \epsilon\int_0^s \frac{S G_\rho h^\epsilon_\sigma G + S G_\rho h^\epsilon_\tau + S G_\sigma G}{G^2}\,dt + \frac{\epsilon^2}{\mu^2}\int_0^s \frac{S S_\tau}{G^2}\,d\tau.\end{aligned}\tag{8.32}$$

Note that, because their integrands are bounded, all terms except the last one on the right-hand side of the final equality of display (8.32) are $O(\epsilon)$. To estimate the last term, let us differentiate the function S^2/G^2 with respect to s, and rearrange the resulting identity, to obtain the following expression:

$$\frac{S S_\tau}{G^2} = \frac{\mu^2}{2}\frac{d}{ds}\Big(\frac{S^2}{G^2}\Big) - \mu^2 S\frac{S_\rho h^\epsilon_\sigma G + S_\rho h^\epsilon_\tau + S_\sigma G}{G^2} \\ + \mu^2 S^2 \frac{G_\rho h^\epsilon_\sigma G + G_\rho h^\epsilon_\tau + G_\sigma G}{G^3} + \epsilon S^2 \frac{S_\tau}{G^3}.$$

If we integrate both sides of the last identity over the interval $[0, s]$, then all the integrands are bounded. In view of this fact and the inequality $\epsilon \le \mu$, it follows that

$$\frac{\epsilon^2}{\mu^2}\int_0^s \frac{S S_\tau}{G^2}\,dt = sO(\epsilon).$$

To estimate the term

$$\int_0^s \frac{G_\rho h^\epsilon_\tau}{G}\,dt$$

that appears in the expression (8.31) for the integral

$$\int_0^s (G_\rho h^\epsilon_\sigma + G_\sigma)\,dt,$$

use the inequality $|h^\epsilon_\tau|_{C^0} \le C\epsilon$ obtained in Lemma 8.3.

In summary, we have $|\int_0^s (G_\rho h^\epsilon_\sigma + G_\sigma)\,dt| \le C_1 + C_2\epsilon s$ for some constants $C_1 > 0$ and $C_2 > 0$ both independent of ϵ. □

LEMMA 8.5. *With the hypotheses of Lemma* 8.3, *if* $\mu > 0$ *is sufficiently small, then there is a constant* $C > 0$ *such that* $|h^\epsilon_{\sigma\sigma}|_{C^0} \le C$.

Proof. The proof of the lemma is similar to the proof of Lemma 8.3.

Recall that h^ϵ_σ satisfies the "fixed point equation" in display (8.23), and choose $\mu > 0$ sufficiently small so that $T(\mu)$ is as in the proof of Lemma 8.3. If we set $p = \sigma(T, q^\epsilon)$ and $s = T$, then we have the identity

$$h^\epsilon_\sigma(p) = \frac{\psi^\epsilon_{11}(T, q^\epsilon) h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{12}(T, q^\epsilon)}{\psi^\epsilon_{21}(T, q^\epsilon) h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{22}(T, q^\epsilon)}. \tag{8.33}$$

By a direct computation of the derivative of both sides of (8.33) with respect to p, we obtain

$$\begin{aligned} h^\epsilon_{\sigma\sigma}(p) = \frac{\frac{dq^\epsilon}{dp}}{(\psi^\epsilon_{21} h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{22})^2} \Bigg(& \left(\psi^\epsilon_{21} \frac{d}{dq} \psi^\epsilon_{11} - \psi^\epsilon_{11} \frac{d}{dq} \psi^\epsilon_{21}\right) (h^\epsilon_\sigma(q^\epsilon))^2 \\ & + \left(\psi^\epsilon_{22} \frac{d}{dq} \psi^\epsilon_{11} - \psi^\epsilon_{12} \frac{d}{dq} \psi^\epsilon_{21} + \psi^\epsilon_{21} \frac{d}{dq} \psi^\epsilon_{12} - \psi^\epsilon_{11} \frac{d}{dq} \psi^\epsilon_{22}\right) h^\epsilon_\sigma(q^\epsilon) \\ & + (\psi^\epsilon_{11} \psi^\epsilon_{22} - \psi^\epsilon_{12} \psi^\epsilon_{21}) h^\epsilon_{\sigma\sigma}(q^\epsilon) + \psi^\epsilon_{22} \frac{d}{dq} \psi^\epsilon_{12} \ - \psi^\epsilon_{12} \frac{d}{dq} \psi^\epsilon_{22} \Bigg), \end{aligned} \tag{8.34}$$

where the functions ψ^ϵ_{ij}, $i, j = 1, 2$, on the right-hand side of the equality are evaluated at (T, q^ϵ).

Using the identity $p = \sigma^\epsilon(T, q^\epsilon)$ and differentiating with respect to p, we have that $dq^\epsilon/dp = 1/\sigma^\epsilon_q(T, q^\epsilon)$. Moreover, using the solution (8.4) and Proposition 8.1, it is not difficult to see that $\sigma^\epsilon_q(T, q) = y^\epsilon(T, q)$, where, y^ϵ is defined in Proposition 8.1. If $\mu > 0$ is sufficiently small so that h^ϵ_σ is sufficiently close to h^0_σ, the matrix $\Psi^\epsilon(T, \cdot)$ is sufficiently close to $\Psi^0(T_0, \cdot)$, and the partial derivative Ψ^ϵ_q is uniformly bounded, then, for all q and all $T = T(\mu)$, we have

$$|\det \Psi^\epsilon(T, q)| < \frac{1}{4} M_0^2, \qquad |(\psi^\epsilon_{21}(T, q) h^\epsilon_\sigma(q) + \psi^\epsilon_{22}(T, q))^2| > \frac{1}{3} M_0,$$

where $M_0 > 0$ is the constant appearing in Lemma 8.3. Thus, using Lemma 8.4, the absolute value of the coefficient

$$\frac{\frac{dq^\epsilon}{dp}}{(\psi^\epsilon_{21} h^\epsilon_\sigma(q^\epsilon) + \psi^\epsilon_{22})^2} (\psi^\epsilon_{11} \psi^\epsilon_{22} - \psi^\epsilon_{12} \psi^\epsilon_{21}) \tag{8.35}$$

will be uniformly bounded less than one.

To estimate the supremum of $h_{\sigma\sigma}$, we proceed in the following order: We take the absolute value of each side of (8.34), apply the triangle law to the right-hand side, take the supremum of the right-hand side over q^ϵ, take the supremum of the left-hand side over p, move the term containing the norm of $h_{\sigma\sigma}$ on the right-hand side to the left-hand side, collect terms, and then divide both sides by the coefficient of the norm of $h_{\sigma\sigma}$. This coefficient is not zero because of the uniform bound on the absolute value of the quantity in display (8.35). Thus, we obtain a uniform bound on the norm of $h_{\sigma\sigma}$. □

9. Proof of Lemma 7.3. If $\mu > 0$ is chosen as in Lemma 8.3 and $\epsilon_* > 0$ is such that, for $0 \le \epsilon < \epsilon_*$, the system $E^{\epsilon,\mu}$ has a k-normally hyperbolic invariant manifold, $k \ge 2$, that is the graph of a C^k function h^ϵ of the angular variables, then, by Lemmas 8.3 and 8.5, the subset $\mathcal{S} := \{h^\epsilon : 0 \le \epsilon < \epsilon_*\}$ in the space of C^2 functions of the angular variables is uniformly bounded. As a result, the set $\mathcal{S}$ is equicontinuous in the C^1 norm. By Arzela's theorem, if we choose a sequence of real numbers increasing to the limit ϵ_*, then we can extract a subsequence $\{\epsilon_k\}$ such that the corresponding sequence $\{h^{\epsilon_k}\}$ converges to a C^1 function h^{ϵ_*}. An easy argument, as in Lemma 5.3, shows that the graph of h^{ϵ_*} is invariant under the flow of $E^{\epsilon_*,\mu}$, as required.

10. Proof of Lemma 7.4. If $\mu > 0$ is chosen as in Lemma 8.3, the number $\epsilon_* > 0$ is such that, for $0 \leq \epsilon < \epsilon_*$, the system $E^{\epsilon,\mu}$ has a k-normally hyperbolic invariant manifold that is the graph of a C^k function h^ϵ of the angular variables, and if $E^{\epsilon_*,\mu}$ has a C^1 invariant manifold $M(\epsilon_*, \mu)$ given as the graph of the function h^{ϵ_*} of the angular variables, then we must show that $M(\epsilon_*, \mu)$ has a continuous invariant normal bundle.

It suffices to construct a normal bundle over the curve

$$q \mapsto (h^{\epsilon_*}(q, 0), q, 0) \tag{10.1}$$

that is invariant with respect to some iterate of the stroboscopic linearized Poincaré map; that is, the map given by moving a point on the slice $\{(\rho, \sigma, \tau) : \tau = 0\}$ forward by the flow to time $2\pi m\mu^2/\Omega$. To prove this reduction, note that the linearized Poincaré map is two-dimensional at each point and that the tangent line to the invariant torus is invariant under the map. Also, for a two-dimensional linear map with an invariant line, if an iterate of the map has two distinct invariant lines, then so does the map. Finally, if there is a normal bundle over the curve, then a normal bundle over the torus is constructed by moving the vectors in the given normal bundle forward by the linearized flow.

We will construct a normal bundle over the curve (10.1). For this, let us consider the function space

$$\Gamma := \{\alpha : \mathbb{S}^1 \to \mathbb{R} : \alpha \in C^0\}.$$

Also, for $\alpha \in \Gamma$, let us define a vector at the point $(h^{\epsilon_*}(q, 0), q, 0)$ as follows:

$$X_0(q) := \alpha(q)X_2(q) + X_2(q)^\perp,$$

where $X_2(q) := \mathcal{X}_2(h^{\epsilon_*}(q, 0), q, 0)$. We will show that there is some choice for $\alpha \in \Gamma$ so that X_0 generates an invariant normal bundle over the curve (10.1).

If n is an integer, $T := 2\pi mn\mu^2/\Omega$, and $p = \sigma^{\epsilon_*}(T, q)$, then X_0 generates an invariant bundle if and only if

$$\lambda^{\epsilon_*}(T, q)X_0(p) = (\alpha(q) + a^{\epsilon_*}(T, q))X_2(p) + b^{\epsilon_*}(T, q)X_2(p)^\perp, \tag{10.2}$$

where λ^{ϵ_*}, a^{ϵ_*}, and b^{ϵ_*} are defined in subsection 8.2 for the system corresponding to ϵ_*. Using these definitions, we find that (10.2) holds if and only if

$$\alpha(p) = \frac{\alpha(q) + a^{\epsilon_*}(T, q)}{b^{\epsilon_*}(T, q)}.$$

Define $T < 0$ analogous to the definition in display (8.24) with the property that, for $0 \leq \epsilon < \epsilon_*$, there is some η such that

$$b^\epsilon(T) := \sup\{b^\epsilon(T, q) : q \in \mathbb{S}^1\} > \eta > 1.$$

Passing to the limit as ϵ approaches ϵ_* from below and using the fact that h^ϵ converges to h^{ϵ_*}, we find that

$$b^{\epsilon_*}(T) := \sup\{b^{\epsilon_*}(T, q) : q \in \mathbb{S}^1\} \geq \eta > 1.$$

Also, let us define $\Lambda : \Gamma \to \Gamma$ by

$$(\Lambda\alpha)(p) = \frac{\alpha(q) + a^{\epsilon_*}(T,q)}{b^{\epsilon_*}(T,q)}.$$

A fixed point of Λ corresponds to the desired invariant normal bundle. But, by a simple computation, we have the inequality

$$|(\Lambda\alpha_2)(p) - (\Lambda\alpha_1)(p)| \le \frac{1}{\eta}|\alpha_2(q) - \alpha_1(q)|.$$

Thus, Λ is a contraction on the complete metric space Γ and Λ has a unique fixed point, as required.

11. Proof of Lemma 7.5. We will show that the C^1 invariant manifold given as the graph of the function h^{ϵ_*} is k-normally hyperbolic under the assumption that this manifold has an invariant normal splitting. For this, we must verify the inequalities given in display (8.11).

Consider $\lambda_3(s)$, and recall formula (8.15). Let us suppose that a bounded neighborhood N as in Proposition 6.2 is chosen, the invariant tori are given by the graphs of the functions h^ϵ of the angular variables, and $\mu > 0$ is sufficiently small so that the invariant manifold given by h^{ϵ_*} is in N. Then, by Lemma 8.3, the functions h^ϵ satisfy the inequality

$$|h^\epsilon - h^0| < C\epsilon \tag{11.1}$$

for $0 \le \epsilon \le \epsilon_*$.

Let $s \to \gamma^{\epsilon_*}(s,q,\tau) = (h^{\epsilon_*}(\sigma^{\epsilon_*}(s,q,\tau), s+\tau), \sigma^{\epsilon_*}(s,q,\tau), s+\tau)$ be the solution of (7.3) corresponding to ϵ_* with the initial condition $(h^{\epsilon_*}(q,\tau), q, \tau)$, let B be the system matrix of the linearization of system (7.3) along the solution γ^ϵ, and note that

$$\operatorname{tr} B(\gamma^{\epsilon_*}(t,q,\tau)) = f_\rho + g_\sigma + \epsilon(R_\rho + S_\sigma).$$

Let ω be the minimal period of the periodic solution of the unperturbed system (7.2), let $(\bar{q}, \bar{\tau})$ be an arbitrary choice of the angular variables, and define

$$b := \int_0^\omega \operatorname{tr} B(\gamma^0(s,\bar{q},\bar{\tau}))\, ds.$$

The quantity b is a Floquet exponent of the periodic orbit that is independent of the choice of the angular variables. By an application of Gronwall's inequality (8.16), there is a constant $C_1 > 0$ such that, for $0 \le s \le \omega$,

$$|\gamma^{\epsilon_*}(s,\bar{q},\bar{\tau}) - \gamma^0(s,\bar{q},\bar{\tau})| \le C_1\epsilon_*.$$

Hence, there is a constant $C_2 > 0$ such that

$$|\operatorname{tr} B(\gamma^{\epsilon_*}(s,\bar{q},\bar{\tau})) - \operatorname{tr} B(\gamma^0(s,\bar{q},\bar{\tau}))| \le C_2\epsilon_*$$

and

$$\int_0^\omega \operatorname{tr} B(\gamma^{\epsilon_*}(s,\bar{q},\bar{\tau}))\, ds \le \int_0^\omega \operatorname{tr} B(\gamma^0(s,\bar{q},\bar{\tau}))\, ds + C_2\epsilon_*\omega = b + C_2\epsilon_*\omega.$$

An arbitrary $s \geq 0$ can be expressed in the form $s = \ell\omega + r$ with $0 \leq r < \omega$. Also, for $k = 1, 2, \dots, \ell$, let us define

$$q_k := \sigma^{\epsilon_*}(k\omega, q, \tau), \quad \tau_k := k\omega + \tau.$$

There are constants $C_3 > 0$ and $C_4 > 0$ such that

$$\begin{aligned}\int_0^s \operatorname{tr} B(\gamma^{\epsilon_*}(t, q, \tau))\, dt =& \sum_{k=0}^{\ell-1} \int_{k\omega}^{(k+1)\omega} \operatorname{tr} B(\gamma^{\epsilon_*}(t, q, \tau))\, dt \\ &+ \int_{\ell\omega}^{\ell\omega + r} \operatorname{tr} B(\gamma^{\epsilon_*}(t, q, \tau))\, dt \\ =& \sum_{k=0}^{\ell-1} \int_0^{\omega} \operatorname{tr} B(\gamma^{\epsilon_*}(t, q_k, \tau_k))\, dt \\ &+ \int_0^{r} \operatorname{tr} B(\gamma^{\epsilon_*}(t, q_\ell, \tau_\ell))\, dt \\ \leq& \ell(b + C_2\epsilon_*\omega) + C_3 \leq \left(\frac{b}{\omega} + C_2\epsilon_*\right) s + C_4.\end{aligned}$$

By Proposition 8.1 and Lemma 8.4, there are constants $C_5 > 0$ and $C_6 > 0$ such that

$$|X_2(s)| \geq C_6 e^{-\mu C_5 s}. \tag{11.2}$$

Also, by Lemma 7.4, the function α corresponding to the normal splitting at ϵ_* is bounded as a periodic function over the invariant manifold.

Taking the above estimates into account and using formula (8.15), there is a constant $c > 0$ such that $\lambda_3(s) \leq ce^{(\frac{b}{\omega} + c\mu)s}$. If, in addition, $\mu > 0$ is sufficiently small, then $-\beta := \frac{b}{\omega} + c\mu < 0$, and we have the desired estimate:

$$\lambda_3(s) \leq ce^{-\beta s} \tag{11.3}$$

for all $s \geq 0$.

The function $|X_1(s)|$ is uniformly bounded below, in fact $|X_1(s)| \geq 1$. Thus, if k is a positive integer, then using the estimate (11.3), there is some $c_1 > 0$ such that

$$\frac{\lambda_3(s)}{\lambda_1^k(s)} \leq c_1 e^{-\beta s}.$$

Using the estimates (11.2) and (11.3), we have that

$$\frac{\lambda_3(s)}{\lambda_1^k(s)} \leq \frac{c_1}{C_6^k} e^{-s(\beta - \mu k c_1)}.$$

Thus, if $\mu > 0$ is sufficiently small, there are constants $c_2 > 0$ and $\beta_1 > 0$ such that

$$\frac{\lambda_3(s)}{\lambda_2^k(s)} \leq c_2 e^{-\beta_1 s}.$$

Finally, using the general smoothness result in [7], it follows that the C^1 manifold given as the graph of h^{ϵ_*} with invariant splitting and with the hyperbolic estimates just proved is in fact a C^k manifold, as required.

12. Appendix.

In this appendix we will prove Propositions 6.1 and 6.2.

Proof of Proposition 6.1. We will construct a family of C^1 curves S_r^+ and S_r^-, $r \in (0,1]$ for system (6.1) such that the following properties are satisfied:

(i) The curves S_r^+ and S_r^- lie in the exterior and interior domains separated by Γ, respectively, and S_r^+ (resp., S_r^-) together with Γ encloses an annulus.

(ii) The curves S_r^+ and S_r^- are transverse to the vector field f.

(iii) There is a constant $C_0 > 0$ that is independent of r such that $\sup\{d(x,\Gamma) : x \in S_r^\pm\} \le C_0 r$.

(iv) If $C_f^\pm(r) := \min_{x\in S_r^\pm}\{\langle f(x), n(x)\rangle\}$, where $n(x)$ is the inward (resp., outward) unit normal vector to S_r^+ (resp., S_r^-) at $x \in S_r^\pm$ and the angle brackets denote the usual inner product, then $C_f^\pm(r) \ge C_1 r$ for some constant $C_1 > 0$ independent of r.

Let us assume for the moment that the above construction is possible and use it to complete the proof of the proposition.

For this, let $N \subset \mathbb{R}^2$ be the annulus such that $\partial N = S_1^+ \cup S_1^-$. If $\|g\|_{C^0}$ is small enough, then there exists an $r_0 \in (0,1]$ such that $\|g\|_{C^0} = C_1 r_0$. Using this fact, we have, for $r > r_0$ and $x \in S_r^\pm$, that

$$\langle f(x) + g(x), n(x)\rangle \ge C_f^\pm(r) - \|g\|_{C^0} \ge C_1 r - C_1 r_0 > 0.$$

Thus, for $r > r_0$, the set $S_r^+ \cup S_r^-$ encloses a positively invariant annulus for the system (6.2) in $\mathbb{R}^2$. It follows that $\bar\Gamma$ is contained in this domain, and, by the estimates given in (iii) and (iv), that $d(x,\Gamma) \le C_0 r_0 = \frac{C_0}{C_1}\|g\|_{C^0}$ for every $x \in \bar\Gamma$, as required.

The proof will be completed by constructing a family of curves that satisfies properties (i)–(iv).

The constructions and the verifications of properties (i)–(iv) for the families S_r^+ and S_r^- are similar. We will give the proof for S_r^+ only. Also, in the arguments to follow we will suppress the superscript "+."

Step 1. Construction of S_r.

Since the periodic solution Γ is asymptotically stable, there exists a neighborhood of Γ, contained in the stable manifold of Γ, with an invariant foliation with respect to the system (6.1) whose leaves are curves. Let $M^s(p)$ denote the leaf through the point $p \in \Gamma$. Also, let $t \mapsto x(t,\xi)$ denote the solution of the differential equation (6.1) with $x(0,\xi) = \xi$, and let $\Phi(t,\xi)$ denote the principal fundamental matrix solution at $t = 0$ of the linearized system along this solution.

Fix a point $q_1 \in M^s(p)$ that lies in the exterior domain separated by Γ, and let $q_0 := x(\omega, q_1)$ be the point where the solution through q_1 first returns to $M^s(p)$. Choose a smooth function $q : [0,1] \to M^s(p) \subset \mathbb{R}^2$ such that the derivative of q, including the left-hand and right-hand derivatives at the end points of its domain, does not vanish, and with the additional properties that $q(0) = q_0$, $q(1) = q_1$, and

$$\dot q(0^+) = \Phi(\omega, q_1)\dot q(1^-). \tag{12.1}$$

The last requirement can be met because, by the invariance of the foliation,

$$\Phi(\omega, q_1) T_{q_1} M^s(p) = T_{q_0} M^s(p).$$

Let $t : [0,1] \to [0,\omega]$ be the linear transformation given by $t(\lambda) = \lambda\omega$, and consider the curve S defined parametrically by $\lambda \mapsto x(t(\lambda), q(\lambda))$. Let us note that S is closed. Indeed, since $t(0) = 0$ and $t(1) = \omega$, we have that $x(t(0), q(0)) = x(t(1), q(1)) = q_0$.

For each $\lambda \in (0,1)$, define $T(\lambda)$ to be the tangent vector to S at the point $x(t(\lambda),q(\lambda))$ given by

$$\begin{aligned} T(\lambda) &= \frac{d}{ds}x(t(s),q(s))\big|_{s=\lambda} = \omega\dot{x}(t(\lambda),q(\lambda)) + x_\xi(t(\lambda),q(\lambda))\dot{q}(\lambda) \\ &= \omega f(x(t(\lambda),q(\lambda))) + \Phi(t(\lambda),q(\lambda))\dot{q}(\lambda). \end{aligned} \tag{12.2}$$

To check that S is a C^1 curve, it suffices to show that $T(0^+) = T(1^-)$. But, this equality follows from the identities (12.1) and (12.2).

Let ϕ^s denote the flow associated with the system (6.1). The family of curves S_r, for $r \in (0,1]$, is defined as follows: $S_r := \phi^s(S)$, where $s = (\omega/b)\ln r$.

Step 2. Verification of properties (i)–(iv) for S_r.

Property (i) is obvious from the construction.

To check property (ii), we will show first that the curve S is transverse to the vector field given by f. Because the vector $\dot{q}(\lambda)$ is tangent to $M^s(p)$ at $q(\lambda)$, this vector is not parallel to the vector $f(q(\lambda))$. Using the fact that

$$\Phi(t(\lambda),q(\lambda))f(q(\lambda)) = f(x(t(\lambda),q(\lambda))),$$

it follows that the vectors $\Phi(t(\lambda),q(\lambda))\dot{q}(\lambda)$ and $f(x(t(\lambda),q(\lambda)))$ are independent at $x(t(\lambda),q(\lambda))$. In view of this fact and the formula (12.2) for the vector $T(\lambda)$ tangent to S, it is clear that f is everywhere transverse to S.

Next, for each point $Q \in S_r$, there exists a point $P \in S$ such that $\phi^s(P) = Q$. Therefore, $T_QS_r = D\phi^s(P)T_PS$ and $f(Q) = D\phi^s(P)f(P)$. Since T_PS is transverse to $f(P)$, we have that T_QS_r is transverse to $f(Q)$. This proves property (ii).

For the proof of property (iii), let us note that, due to the hyperbolicity of the orbit Γ, there exists some $C_0 > 0$ such that, for each point $x_0 \in N$, we have $d(\phi^s(x_0),\Gamma) \le C_0e^{bs/\omega}$. Hence, for each $Q \in S_r$, if we take the point $P \in S$ such that $\phi^s(P) = Q$, then

$$d(Q,\Gamma) = d(\phi^s(P),\Gamma) \le C_0e^{\frac{bs}{\omega}} = C_0r,$$

and the constant C_0 depends only on the "size" of the neighborhood N.

Finally, let us prove property (iv). To this end, note that for each point $Q \in S_r$, there is a corresponding point $P \in S$ such that $Q = \phi^s(P)$ and some λ such that $P = x(t(\lambda),q(\lambda))$. Also, with an abuse of notation, let $T(Q)$ denote the vector in T_QS_r given by $T(Q) = D\phi^s(P)T(\lambda)$, and define $n(Q)$ to be the inward unit normal vector to S_r at Q.

Using the easily verified identity $\langle f(Q),n(Q)\rangle|T(Q)| = |f(Q)\times T(Q)|$, let us note that if $r \in (0,1]$, then

$$C_f^+(r) = \min_{Q\in S_r}\left\{\frac{|f(Q)\times T(Q)|}{|T(Q)|}\right\}.$$

Also, recall formula (12.2), and note that

$$\begin{aligned} T(Q) = D\phi^s(P)T(\lambda) &= \omega\Phi(s,P)f(P) + \Phi(s,P)\Phi(t(\lambda),q(\lambda))\dot{q}(\lambda) \\ &= \omega f(Q) + \Phi(s+t(\lambda),q(\lambda))\dot{q}(\lambda). \end{aligned}$$

Moreover, we have that $|f(Q)\times T(Q)| = |f(Q)\times\Phi(s+t(\lambda),q(\lambda))\dot{q}(\lambda)|$.

By an initial choice of the point q_1 sufficiently close to $p \in M^s(p)$, there is a number $K \geq 1$ such that

$$\frac{1}{K}\exp\left(\int_0^{s+t(\lambda)} \operatorname{tr} A(\tau)\, d\tau\right) \leq |\Phi(s+t(\lambda), q(\lambda))\dot{q}(\lambda)| \leq K\exp\left(\int_0^{s+t(\lambda)} \operatorname{tr} A(\tau)\, d\tau\right),$$

where $A(t) = Df(x(t,p))$.

Let $v(Q)$ denote the unit tangent vector at Q to the stable fiber through Q. Because $\Phi(s+t(\lambda), q(\lambda))\dot{q}(\lambda)$ is tangent to the stable fiber through Q, we have that

$$\begin{aligned}|f(Q) \times T(Q)| &= |f(Q) \times \Phi(s+t(\lambda), q(\lambda))\dot{q}(\lambda)|\\ &\geq \frac{1}{K}|f(Q) \times v(Q)| \exp\left(\int_0^{s+t(\lambda)} \operatorname{tr} A(\tau)\, d\tau\right),\end{aligned}$$

and there is a constant $C_2 > 0$ such that

$$\begin{aligned}|T(Q)| &\leq \omega|f(Q)| + |\Phi(s+t(\lambda), q(\lambda))\dot{q}(\lambda)|\\ &\leq \omega|f(Q)| + K\exp\left(\int_0^{s+t(\lambda)} \operatorname{tr} A(\tau)\, d\tau\right)\\ &\leq \omega|f(Q)| + C_2.\end{aligned}$$

Therefore, using the above estimates and the fact that the quantities $|f|$ and $|f(Q) \times v(Q)|$ are bounded below over N, there is a constant $C_3 > 0$ such that

$$\begin{aligned}\frac{|f(Q) \times T(Q)|}{|T(Q)|} &\geq \frac{|f(Q) \times v(Q)|}{K(\omega|f(Q)| + C_2)} \exp\left(\int_0^{s+t(\lambda)} \operatorname{tr} A(\tau)\, d\tau\right)\\ &\geq C_3 \exp\left(\int_0^{s} \operatorname{tr} A(\tau)\, d\tau\right).\end{aligned}$$

If m is a nonnegative integer and $0 \leq \lambda < \omega$ is such that $s = m\omega + \lambda$, then there are constants $C_1 > 0$ and $C_4 > 0$ such that

$$\begin{aligned}C_3 \exp\left(\int_0^{s} \operatorname{tr} A(\tau)\, d\tau\right) &= C_3 \exp\left(\int_0^{m\omega+\lambda} \operatorname{tr} A(\tau)\, d\tau\right)\\ &\geq C_4 \exp\left(\int_0^{m\omega} \operatorname{tr} A(\tau)\, d\tau\right)\\ &= C_4 e^{bm} = C_4 e^{\frac{bs-b\lambda}{\omega}} \geq C_1 r.\end{aligned}$$

Thus, we have proved that $C_f(r) \geq C_1 r$. □

Proof of Proposition 6.2. If the families of curves S_r^+ and S_r^- are "suspended" to tori in the space $\mathbb{R}^2 \times \mathbb{R}$, then the corresponding tori can be shown to satisfy conditions analogous to the conditions (i)–(iv) that are defined in the proof of Proposition 6.1. The verification of each condition is essentially identical to the corresponding proof in Proposition 6.1; we omit the details. □

Acknowledgments. Finally, we thank the referees for carefully reading the original version of this paper. Their comments led to many improvements.

REFERENCES

[1] N. N. Bogoliubov and Y. A. Mitropolsky, *The method of integral manifolds in nonlinear mechanics*, Contrib. Differential Equations, 2 (1963), pp. 123–196.

[2] C. Chicone, *Invariant tori for periodically perturbed oscillators*, Publ. Mat., 41 (1997), pp. 57–83.

[3] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.

[4] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[5] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* 2nd ed., Springer-Verlag, New York, 1986.

[6] J. Hale, *Ordinary Differential Equations,* Robert E. Krieger Pub. Co. Inc., Huntington, NY, 1980.

[7] M. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math., vol. 583, Springer-Verlag, New York, 1977.

[8] N. Kopell, *Invariant manifolds and the initialization problem for some atmospheric equations*, Phys. D, 14 (1985), pp. 203–215.

[9] V. K. Melnikov, *On the stability of the center for time-periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1–56.

[10] C. Robinson and J. Murdock, *Some mathematical aspects of spin-orbit resonance*. II. Celestial Mech., 24 (1981), 83–107.

[11] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos,* Springer-Verlag, New York, 1990.

[12] S. Wiggins, *Normally Hyperbolic Invariant Manifolds in Dynamical Systems*, Springer-Verlag, New York, 1994.

SIAM J. Math. Anal.
Vol. 31, No. 2, pp. 416–430

DIRECTIONAL AND TIME-SCALE WAVELET ANALYSIS[*]

ROB A. ZUIDWIJK[†]

Abstract. Combined use of the X-ray (Radon) transform and the wavelet transform has proved to be useful in application areas such as diagnostic medicine and seismology. The wavelet X-ray transform performs one-dimensional wavelet transforms along lines in $\mathbb{R}^n$ which are parameterized in the same fashion as for the X-ray transform. The reconstruction formula for this transform gives rise to a continuous family of elementary projections. These projections provide the building blocks of a directional wavelet analysis of functions in several variables. Discrete wavelet X-ray transforms are described which make use of wavelet orthonormal bases and, more generally, of biorthogonal systems of wavelet Riesz bases. Some attention is given to approximation results which involve wavelet X-ray analysis in several directions.

Key words. wavelet X-ray transform, wavelet transform, X-ray transform, Radon transform, windowed Radon transform, local Radon transform, reconstruction formula, wavelet orthonormal basis, biorthogonal wavelet expansion, wavelet frame

AMS subject classifications. 42C15, 44A12, 86A15

PII. S0036141098333359

1. Introduction and preliminaries.

The wavelet transform and the X-ray transform (or Radon transform) and their discretizations have received considerable attention in the mathematical literature. Moreover, the two transforms have proved to be very useful as a tool to handle a variety of engineering problems. See references [Chu, Dau, Hol, Mey, RV] for the wavelet transform and [Dea, Nat, Sol, SSW] for the X-ray transform.

A combined use of the two transforms has also proved to be useful. Among the application areas are seismology and diagnostic medicine. Indeed, localized inversion of the Radon transform using wavelets [BW, OD] can be applied in diagnostic medicine to reduce the radiation exposure when X-rays of only a small area in a local region of tissue are required. This technique can also be applied to cross-borehole tomography in seismic exploration; see [DL].

On the other hand, the Radon transform has been applied extensively in reflection seismology. If one models the earth's subsurface as a stratified medium, then the Radon transform can be used to transform seismic data in such a way that arriving wavefronts with distinct propagating velocities are separated. In this context, the Radon transform is referred to as a *slant stack* [Rob].

We shall now shortly describe the aforementioned integral transforms. The *X-ray transform*

$$Pf(\theta, x) = \int_{\mathbb{R}} f(x + t\theta)\, dt$$

integrates a function f on $\mathbb{R}^n$ along an affine line $x+\mathbb{R}\theta$, where $x \in \mathbb{R}^n$ is perpendicular to the direction θ. Observe that (θ, x), where θ is a vector on the unit sphere $S^{n-1} = \{y \in \mathbb{R}^n \mid \|y\| = 1\}$ and x a vector orthogonal to θ, parameterize all lines in $\mathbb{R}^n$. In

[*] Received by the editors January 29, 1998; accepted for publication (in revised form) April 20, 1999; published electronically January 10, 2000. This work was supported by the Technology Foundation (STW), project CWI44.3403.

http://www.siam.org/journals/sima/31-2/33335.html

[†] Rotterdam School of Management, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (R.Zuidwijk@fac.fbk.eur.nl).

particular, the distance of the line $x + \mathbb{R}\theta$ to the origin is given by $\|x\|$. The relevance of the X-ray transform to diagnostic medicine can be understood as follows. The attenuation of X-ray beams (along lines in $\mathbb{R}^2$ or $\mathbb{R}^3$) passing through a medium with density f is modeled by the integral of the density function along these lines. It is the aim of computerized tomography to reconstruct the density function from these attenuation data, i.e., from the X-ray transformed function [SSW, Nat].

The *continuous wavelet transform*

$$W_g f(b,a) = \int_{\mathbb{R}} f(t) \frac{1}{\sqrt{a}} \overline{g\left(\frac{t-b}{a}\right)}\, dt, \quad b \in \mathbb{R}, \quad a > 0,$$

which puts a function f to its wavelet coefficients $W_g f(b,a)$, is often considered as an alternative for the windowed Fourier transform in the time-frequency analysis of transient signals; e.g., see [Mey, RV, Wal]. The transform actually computes inner products of f with respect to translated and dilated versions of one and the same function g, which is referred to as the wavelet. Usually, the function g satisfies an admissibility condition to ensure that the function f can be reconstructed from its wavelet coefficients $W_g f$; details are given in section 2.

In [FKV, FKV2], it has been argued that the Radon transform (as a slant stack) and the wavelet transform (as a time-frequency analysis tool) have complementary useful features to remove noise from seismic reflection data. For this reason, the two transforms are applied in a cascaded fashion. This work motivated the definition of a transformation which combines the properties of the wavelet and the X-ray transform. Indeed, we consider the *wavelet X-ray transform*

$$P_g f(\theta, x, b, a) = \int_{\mathbb{R}} f(x + t\theta) \frac{1}{\sqrt{a}} \overline{g\left(\frac{t-b}{a}\right)}\, dt.$$

This transform computes one-dimensional wavelet transforms along lines in $\mathbb{R}^n$ which are parameterized in the same fashion as for the X-ray transform.

Starting from the reconstruction formulas for the wavelet X-ray transform given in section 3, it will be shown in section 4 that a function $f \in L^2(\mathbb{R}^n)$ can be analyzed into elementary projections $G_{\theta,b,a} f \in L^2(\mathbb{R}^n)$ which are parameterized by direction $\theta \in S^{n-1}$, position (or time) $b \in \mathbb{R}$, and scale $a > 0$. In other words, a directional wavelet analysis is performed on functions in several variables.

An alternative approach towards directional wavelet analysis in two or more dimensions uses a continuous wavelet transform on functions in two or more variables incorporating rotation, translation, and dilation [AM]. In this case, wavelets in two or more variables are not only translated and dilated but also rotated.

The wavelet X-ray transform originates from [KS], where it was called the windowed Radon transform. In that paper, the theory continues into the direction of the analytic signal transform. Reconstruction formulas for the wavelet X-ray transform given there (see also [Tak]) are improved in this paper; see Theorem 3.2.

Further, as a localized Radon transform, the wavelet X-ray transform has been used to detect linear events in radar images [WD].

The setup of this paper reads as follows. In sections 2 and 3, respectively, the continuous wavelet transform and the continuous wavelet X-ray transform are discussed briefly. In section 4, the notion of an elementary projection $G_{\theta,b,a} f$ of a function $f \in L^2(\mathbb{R}^n)$ is introduced. It is shown that the operators $G_{\theta,b,a} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ form a continuous family of projection operators; see Theorem 4.6.

In section 5, the continuous wavelet X-ray transform is discretized using an orthonormal basis of wavelets in $L^2(\mathbb{R})$. It is shown in Theorem 5.5 that any function $f \in L^2(\mathbb{R})$ can be written as a series of elementary projections with fixed direction $\theta \in S^{n-1}$. Therefore, each function can be approximated with arbitrary precision by a finite sum of elementary projections. Proposition 5.7 shows that the use of several directions improves the performance of the approximation; see Lemma 5.8. In section 6, the discretization of the wavelet X-ray transform is carried out using biorthogonal systems of Riesz bases. Part of the results there are proved for frame systems.

Some remarks on notation in this paper are in order. The inner product and induced norm on an infinite-dimensional Hilbert space H are denoted by $\langle\cdot,\cdot\rangle_H$ and $\|\cdot\|_H$. The inner product and the Euclidian norm on $\mathbb{R}^n$, however, are indicated by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. It should not lead to confusion that the (spectral) norm of a linear operator $A : H_1 \to H_2$ between Hilbert spaces is also indicated by $\|A\|$. Further, we write $\ker A = \{x \in H_1 \mid Ax = 0\}$ and $\operatorname{ran} A = \{Ax \mid x \in H_1\}$. The open upper half-plane is denoted by $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}$ and the unit sphere in $\mathbb{R}^n$ is denoted by $S^{n-1} = \{z \in \mathbb{R}^n \mid \|z\| = 1\}$. We shall denote the measure $db\, a^{-2}da$ on the open upper half-plane by $d\mu(b, a)$.

In the last part of this section, some preliminary material concerning Riesz bases and frames will be presented; for more information, we refer to [You] and to relevant parts in the wavelet literature; see, for example, [CR, Chu, Dau, Hol]. Recall that a *Riesz basis* in a Hilbert space H is a sequence of vectors $(x_k)_{k=1}^{\infty}$ with closed linear span equal to H and with constants $0 < A \le B$ such that the following holds: for each finite sequence $(a_k)_{k=1}^{N}$, we have

$$A\sum_{k=1}^{N}|a_k|^2 \le \left\|\sum_{k=1}^{N} a_k x_k\right\|_H^2 \le B\sum_{k=1}^{N}|a_k|^2.$$

If the constants A, B are chosen optimally, then they are referred to as the *Riesz bounds* of $(x_k)_{k=1}^{\infty}$. We shall consider Riesz bases $(x_k)_{k=1}^{\infty}$ and $(\widetilde{x}_k)_{k=1}^{\infty}$ which satisfy the *biorthogonality condition*

$$\langle x_k, \widetilde{x}_l\rangle_H = \begin{cases} 1, & k = l, \\ 0, & k \ne l. \end{cases} \tag{1.1}$$

Let $(x_k)_{k=1}^{\infty}$ be a Riesz basis in a Hilbert space H with Riesz bounds A, B. There exists a unique Riesz basis $(\widetilde{x}_k)_{k=1}^{\infty}$ in H such that the biorthogonality condition (1.1) is satisfied. Moreover, the Riesz bounds of $(\widetilde{x}_k)_{k=1}^{\infty}$ are given by $\widetilde{A} = B^{-1}, \widetilde{B} = A^{-1}$.

A *frame* $(x_k)_{k=1}^{\infty}$ in a Hilbert space H is a sequence of vectors with constants $0 < A \le B$ such that for all $y \in H$, we get

$$A\|y\|_H^2 \le \sum_{k=1}^{\infty}|\langle y, x_k\rangle_H|^2 \le B\|y\|_H^2.$$

If the constants A, B are chosen optimally, then they are called the *frame bounds* of $(x_k)_{k=1}^{\infty}$. Observe that if $y \perp x_k$ for all $k \in \mathbb{Z}^+$, then $y = 0$. This implies that the closed linear span of the frame vectors is the whole Hilbert space H. However, the frame vectors need not be linearly independent. We also mention that each Riesz basis is a frame. In general, a biorthogonality condition for frames does not make sense because of lack of linear independence. Nevertheless, expansions with frames are possible using a so-called dual frame. Indeed, consider the operator $F : H \to \ell^2(\mathbb{Z}^+)$, given by

$$Fy = (\langle y, x_k \rangle_H)_{k=1}^{\infty}, \quad y \in H.$$

By construction, F is a bounded operator on H. Its adjoint $F^* : \ell^2(\mathbb{Z}^+) \to H$ is given by

$$F^*(a_k)_{k=1}^{\infty} = \sum_{k=1}^{\infty} a_k x_k, \quad \sum_{k=1}^{\infty} |a_k|^2 < \infty.$$

The operator $F^*F : H \to H$, called the *frame operator*, is a positive boundedly invertible operator; in fact, $A \cdot I_H \le F^*F \le B \cdot I_H$. Define $\widetilde{x}_k = (F^*F)^{-1}x_k$ for $k \in \mathbb{Z}^+$. It turns out that the sequence of vectors $(\widetilde{x}_k)_{k=1}^{\infty}$ is a frame, and it is called the *dual frame* of $(x_k)_{k=1}^{\infty}$. By construction,

$$y = \sum_{k=1}^{\infty} \langle y, \widetilde{x}_k \rangle_H x_k = \sum_{k=1}^{\infty} \langle y, x_k \rangle_H \widetilde{x}_k, \quad y \in H.$$

The frame $(\widetilde{x}_k)_{k=1}^{\infty}$ has frame bounds $\widetilde{A} = B^{-1}$ and $\widetilde{B} = A^{-1}$.

2. Continuous wavelet transform. In this section, we discuss the continuous wavelet transform. We shall consider the general case when the analyzing wavelet need not equal the reconstructing wavelet. Results in this section are without proof. The proofs can be found in, e.g., [Hol, Koo]. For $f, g \in L^2(\mathbb{R})$, consider the expression

$$W_g f(b,a) = \int_{\mathbb{R}} f(t) \overline{\frac{1}{\sqrt{a}} g\left(\frac{t-b}{a}\right)} \, dt, \quad (b,a) \in \mathbb{H},$$

where $\mathbb{H} = \{(b,a) \in \mathbb{R}^2 \mid a > 0\}$ denotes the open upper half-plane. We shall refer to W_g as the *continuous wavelet transform.* The function g plays the role of the *analyzing wavelet* and will be accompanied by a *reconstructing wavelet* $h \in L^2(\mathbb{R})$. The pair of wavelets g, h normally satisfies an admissibility condition which will be specified later. Introducing the shorthand notation

$$g_{b,a}(t) = \frac{1}{\sqrt{a}} g\left(\frac{t-b}{a}\right), \quad g_a(t) = g_{0,a}(t), \quad t \in \mathbb{R}, \quad (b,a) \in \mathbb{H},$$

we get $W_g f(b,a) = \langle f, g_{b,a} \rangle_{L^2(\mathbb{R})}$. Observe that $g_{b,a}$ represents a dilated and translated version of the function g, which is normalized in such a way that $\|g_{b,a}\|_{L^2(\mathbb{R})} = \|g\|_{L^2(\mathbb{R})}$. The Fourier transform $f \mapsto \widehat{f}$, given by

$$\widehat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-i\omega x} \, dx,$$

is a unitary operator on $L^2(\mathbb{R})$. In the following theorem, which states Parseval's formula for the continuous wavelet transform, the pair of wavelets $g, h \in L^2(\mathbb{R})$ should satisfy the following *admissibility condition*: the expression

$$c_{g,h} = \int_0^{\infty} \frac{\overline{\widehat{g}(a\omega)} \widehat{h}(a\omega)}{a} \, da$$

must assume a nonzero value and should be constant for almost all $\omega \in \mathbb{R}$. A pair of wavelets $g, h \in L^2(\mathbb{R})$ which satisfies this admissibility condition will be called an *admissible pair of wavelets*. A wavelet $g \in L^2(\mathbb{R})$ is called an *admissible wavelet* if the pair g, g is an admissible pair of wavelets.

THEOREM 2.1 (Parseval's formula). *Let $g, h \in L^2(\mathbb{R})$ be an admissible pair of wavelets; then for $f, k \in L^2(\mathbb{R})$, one gets*

$$\langle W_g f, W_h k\rangle_{L^2(\mathbb{H})} = c_{g,h} \cdot \langle f, k\rangle_{L^2(\mathbb{R})}.$$

We now proceed with a reconstruction formula for the continuous wavelet transform. Observe that, in this theorem, we require the admissible pair of wavelets to consist of wavelets which are admissible themselves.

THEOREM 2.2 (reconstruction formula). *Let $g, h \in L^2(\mathbb{R})$ be an admissible pair of admissible wavelets, and assume that $f \in L^2(\mathbb{R})$. Then*

$$f = \frac{1}{c_{g,h}} \int_{\mathbb{H}} W_g f(b,a) h_{b,a}(\cdot)\, d\mu(b,a).$$

The integral converges in the norm of $L^2(\mathbb{R})$.

3. Continuous wavelet X-ray transform. We shall now consider a transform acting on square integrable functions on $\mathbb{R}^n$. This transform actually performs one-dimensional wavelet transforms (see the preceding section) along lines in $\mathbb{R}^n$. These lines are parameterized in the same fashion as for the usual X-ray transform (see [Nat, Sol]), i.e., by means of the vector bundle on the unit sphere

$$\mathcal{T} = \{(\theta, x) \mid \theta \in S^{n-1}, x \in \theta^\perp\}.$$

Here $\theta^\perp$ denotes the orthogonal complement of $\theta \in S^{n-1}$ in $\mathbb{R}^n$. Let $g \in L^2(\mathbb{R})$, $f \in L^2(\mathbb{R}^n)$, and define

$$P_g f(\theta, x, b, a) = \int_{\mathbb{R}} f(x + t\theta)\overline{g_{b,a}(t)}\, dt, \quad (\theta, x) \in \mathcal{T}, \quad (b,a) \in \mathbb{H}.$$

The transform P_g will be called the *continuous wavelet X-ray transform.* If we fix $\theta \in S^{n-1}$, we shall write

$$P_{g,\theta} f(x, b, a) = P_g(\theta, x, b, a), \qquad x \in \theta^\perp,\ (b,a) \in \mathbb{H}.$$

We shall also formulate results in terms of this transform, i.e., for the wavelet X-ray transform with fixed direction $\theta \in S^{n-1}$. In the next theorem, we derive Parseval's formulas for the wavelet X-ray transforms. The proofs of the following two theorems can be found in [Zui, Zui2] for the case when $g = h$. For completeness, we shall give an outline of the proof of Theorem 3.2.

THEOREM 3.1 (Parseval's formulas). *Let $g, h \in L^2(\mathbb{R})$ be an admissible pair of admissible wavelets and assume that $f, k \in L^2(\mathbb{R}^n)$. Then*

$$\langle P_g f, P_h k\rangle_{L^2(\mathcal{T}\times\mathbb{H})} = c_{g,h} \cdot |S^{n-1}| \cdot \langle f, k\rangle_{L^2(\mathbb{R}^n)}.$$

Moreover, for fixed $\theta \in S^{n-1}$, we get

$$\langle P_{g,\theta} f, P_{h,\theta} k\rangle_{L^2(\theta^\perp\times\mathbb{H})} = c_{g,h} \cdot \langle f, k\rangle_{L^2(\mathbb{R}^n)}.$$

THEOREM 3.2. *Let $g, h \in L^2(\mathbb{R})$ be an admissible pair of admissible wavelets; then for any $f \in L^2(\mathbb{R}^n)$, one gets*

$$f = \frac{1}{c_{g,h} \cdot |S^{n-1}|} \int_{S^{n-1}} \int_{\mathbb{H}} P_g f(\theta, E_\theta \cdot, b, a) h_{b,a}(\langle \cdot, \theta \rangle) \, d\mu(b,a) \, d\theta, \tag{3.1}$$

where $E_\theta = I - \langle \cdot, \theta \rangle \theta$ denotes the orthoprojector onto $\theta^\perp \subseteq \mathbb{R}^n$. Moreover, for $\theta \in S^{n-1}$ fixed,

$$f = \frac{1}{c_{g,h}} \int_{\mathbb{H}} P_{g,\theta} f(E_\theta \cdot, b, a) h_{b,a}(\langle \cdot, \theta \rangle) \, d\mu(b,a). \tag{3.2}$$

Both integrals converge in $L^2(\mathbb{R}^n)$.

Proof. We will prove (3.2). In the main part of the proof we will replace $P_{g,\theta} f$ by an arbitrary $\Sigma \in L^2(\theta^\perp \times \mathbb{H})$. Let $(K_m)_{m=1}^\infty$ denote an increasing sequence of compact subsets in the open upper half-plane such that $\bigcup_{m=1}^\infty K_m = \mathbb{H}$. For fixed $m \in \mathbb{Z}^+$, consider the expression

$$\mathcal{I}_m = \frac{1}{c_{g,h}} \int_{K_m} \Sigma(E_\theta \cdot, b, a) h_{b,a}(\langle \cdot, \theta \rangle) \, d\mu(b,a).$$

First, one needs to show that $\mathcal{I}_m \in L^2(\mathbb{R}^n)$. This follows from

$$\int_{\mathbb{R}^n} |\mathcal{I}_m(y)|^2 \, dy = \int_{\theta^\perp} \int_{\mathbb{R}} \frac{1}{|c_{g,h}|^2} \left| \int_{K_m} \Sigma(x,b,a) h_{b,a}(t) \, d\mu(b,a) \right|^2 dt \, dx$$

$$\leq \frac{\mu(K_m)}{|c_{g,h}|^2} \|h\|^2_{L^2(\mathbb{R})} \int_{\theta^\perp} \int_{K_m} |\Sigma(x,b,a)|^2 \, d\mu(b,a) \, dx \leq \frac{\mu(K_m)}{|c_{g,h}|^2} \|h\|^2_{L^2(\mathbb{R})} \|\Sigma\|^2_{L^2(\theta^\perp \times \mathbb{H})}.$$

Convergence of the sequence $(\mathcal{I}_m)_{m=1}^\infty$ in $L^2(\mathbb{R}^n)$ follows from the following estimate. Let $p < q$ be positive integers and $k \in L^2(\mathbb{R}^n)$; then application of Fubini's theorem gives

$$\left| \langle \mathcal{I}_q - \mathcal{I}_p, k \rangle_{L^2(\mathbb{R}^n)} \right| = \left| \frac{1}{c_{g,h}} \int_{K_q \setminus K_p} \int_{\theta^\perp} \Sigma(x,b,a) \overline{P_{h,\theta} k(x,b,a)} \, dx \, d\mu(b,a) \right|$$

$$\leq \frac{\sqrt{c_{h,h}}}{|c_{g,h}|} \cdot \sqrt{\int_{K_q \setminus K_p} \int_{\theta^\perp} |\Sigma(x,b,a)|^2 \, dx \, d\mu(b,a)} \cdot \|k\|_{L^2(\mathbb{R}^n)} \to 0$$

as $p, q \to \infty$. Finally, after the substitution $\Sigma = P_{g,\theta} f$, we get

$$\lim_{m \to \infty} \langle \mathcal{I}_m, k \rangle_{L^2(\mathbb{R}^n)} = \lim_{m \to \infty} \frac{1}{c_{g,h}} \langle P_{g,\theta} f, P_{h,\theta} k \rangle_{L^2(\theta^\perp \times K_m)} = \langle f, k \rangle_{L^2(\mathbb{R}^n)}.$$

This proves (3.2). Formula (3.1) now follows trivially. □

In the remainder of this paper, we shall fix $\theta \in S^{n-1}$ in many places, but we will always use $P_g(\theta, x, b, a)$ instead of $P_{g,\theta}(b,a)$.

4. Elementary projections. In this section, we assume that $g, h \in L^2(\mathbb{R})$ is an admissible pair of admissible wavelets. The normalization $\langle g, h\rangle_{L^2(\mathbb{R}^n)} = 1$ is required for the statement in Lemma 4.1. This lemma shows that the integrands in the right-hand sides of (3.1) and (3.2) give rise to the definition of projections on $L^2(\mathbb{R}^n)$.

LEMMA 4.1. *The linear mapping $G_{\theta,b,a} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ given by*

$$G_{\theta,b,a}f(y) = P_g f(\theta, E_\theta y, b, a)\, h_{b,a}(\langle y, \theta\rangle), \quad y \in \mathbb{R}^n, \tag{4.1}$$

is a projection with norm $\|G_{\theta,b,a}\| = \|g\|_{L^2(\mathbb{R})} \cdot \|h\|_{L^2(\mathbb{R})}$.

We shall call $G_{\theta,b,a}$ an *elementary projection* for each direction $\theta \in S^{n-1}$ and each translation-dilation pair $(b, a) \in \mathbb{H}$. Observe that the projection $G_{\theta,b,a}$ is an orthoprojector if and only if $g = h$.

Proof. We first prove that $G_{\theta,b,a}$ is a bounded operator. Indeed,

$$\|G_{\theta,b,a}f\|^2_{L^2(\mathbb{R}^n)} = \int_{\theta^\perp}\int_{\mathbb{R}} |P_g f(\theta, x, b, a)|^2 |h_{b,a}(t)|^2\, dt\, dx$$

$$= \|h\|^2_{L^2(\mathbb{R})} \cdot \int_{\theta^\perp} |P_g f(\theta, x, b, a)|^2\, dx \le \|g\|^2_{L^2(\mathbb{R})} \cdot \|h\|^2_{L^2(\mathbb{R})} \cdot \|f\|^2_{L^2(\mathbb{R}^n)}.$$

We claim that $\|G_{\theta,b,a}\| = \|g\|_{L^2(\mathbb{R})} \cdot \|h\|_{L^2(\mathbb{R})}$. This can be seen by considering $f = \rho(E_\theta \cdot) g_{b,a}(\langle \cdot, \theta\rangle)$ for some $\rho \in L^2(\theta^\perp)$. It is rather straightforward to verify that $G_{\theta,b,a}$ is idempotent, i.e., $G_{\theta,b,a} = G^2_{\theta,b,a}$. □

The following lemmas will be used to prove Theorem 4.6. We omit their straightforward proofs.

LEMMA 4.2. *Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible linear mapping; then $D_A : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$, defined by $D_A f = \sqrt{|\det A|} f(A\cdot)$, is unitary. Moreover, if $f \in L^2(\mathbb{R}^n)$, then $D_A f \to f$ as $A \to I_n$.*

LEMMA 4.3. *For $h \in \mathbb{R}^n$, the linear operator $T_h : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ given by $T_h f = f(\cdot + h)$ is unitary. Moreover, for each $f \in L^2(\mathbb{R}^n)$, we get $T_h f \to f$ as $h \to 0$.*

LEMMA 4.4. *Let $U : \mathbb{R}^n \to \mathbb{R}^n$ be an orthogonal linear mapping. Then*

$$G_{U\theta,b,a} = D_U^{-1} G_{\theta,b,a} D_U.$$

LEMMA 4.5. *Given $(b, a) \in \mathbb{H}$ and $\theta \in S^{n-1}$, we get*

$$G_{\theta,0,1} = D_{aI_n} T_{b\theta} G_{\theta,b,a} T_{b\theta}^{-1} D_{aI_n}^{-1}.$$

We shall prove in Theorem 4.6 that the elementary projections $G_{\theta,b,a}$ depend continuously on the parameters $\theta \in S^{n-1}$ and $(b, a) \in \mathbb{H}$ in the following sense: fix $f \in L^2(\mathbb{R}^n)$; then $G_{\varphi,\beta,\alpha}f \to G_{\theta,b,a}f$ whenever $\varphi \to \theta$ and $(\beta, \alpha) \to (b, a)$.

THEOREM 4.6. *Fix $f \in L^2(\mathbb{R}^n)$; then*

$$G_{\varphi,\beta,\alpha}f \to G_{\theta,b,a}f$$

in $L^2(\mathbb{R}^n)$ whenever $\varphi \to \theta$ and $(\beta, \alpha) \to (b, a)$ in the topologies of $\mathbb{R}^n$ and $\mathbb{H}$, respectively.

Proof. By Lemma 4.5, we may and do assume that $(b, a) = (0, 1)$. Moreover, let U be a rotation in $\mathbb{R}^n$ such that $U\theta = \varphi$ and $\|I_n - U\| = \|\theta - \varphi\|$.

Write $V = D_{\alpha I_n} T_{\beta\theta} D_U$. By Lemmas 4.4 and 4.5,

$$G_{\varphi,\beta,\alpha}f - G_{\theta,0,1}f = (I - V)G_{\varphi,\beta,\alpha}f - G_{\theta,0,1}(I - V)f.$$

Therefore,

$$\|G_{\varphi,\beta,\alpha}f - G_{\theta,0,1}f\|_{L^2(\mathbb{R}^n)} \leq \|(I-V)G_{\varphi,\beta,\alpha}f\|_{L^2(\mathbb{R}^n)} + \|(I-V)f\|_{L^2(\mathbb{R}^n)}.$$

Given $F \in L^2(\mathbb{R}^n)$, we get $(I-V)F = (I-D_{\alpha I_n})F + D_{\alpha I_n}(I-T_{\beta\theta})F + D_{\alpha I_n}T_{\beta\theta}(I-D_U)F$; so

$$\|(I-V)F\|_{L^2(\mathbb{R}^n)} \leq \|(I-D_{\alpha I_n})F\|_{L^2(\mathbb{R}^n)} + \|(I-T_{\beta\theta})F\|_{L^2(\mathbb{R}^n)} + \|(I-D_U)F\|_{L^2(\mathbb{R}^n)} \to 0$$

as $\beta \to 0$, $\alpha \to 1$, and $U \to I_n$, by Lemmas 4.2 and 4.3. This proves the theorem. □

5. Discrete wavelet X-ray transform: Orthonormal case. In this section, we study a discretization of the wavelet X-ray transform, based on the existence of wavelet orthonormal bases. We will assume that for the wavelet $g \in L^2(\mathbb{R}^n)$, there exists a countable set $K \in \mathbb{H}$ such that $(g_{b,a})_{(b,a)\in K}$ defines an orthonormal basis in $L^2(\mathbb{R})$. As an example of such a basis, we mention the case when $K = \{(k2^j, 2^j) \mid k, j \in \mathbb{Z}\}$ and where g is a Daubechies wavelet of a certain order [Dau, Dau2]. Other examples are Battle–Lemarié wavelets and Meyer wavelets; see [CR] and [Dau]. We shall consider the previous results on elementary projections for the special case when $g = h$, i.e., the case when $G_{\theta,b,a}$ is an orthogonal projection. In the next section, the general case will be dealt with. To avoid confusion, we shall write $G^o_{\theta,b,a}$ in the particular case when $g = h$, i.e.,

$$G^o_{\theta,b,a}f = P_g f(\theta, E_\theta\cdot, b, a)g_{b,a}(\langle\cdot,\theta\rangle), \quad \theta \in S^{n-1}, \quad (b,a) \in \mathbb{H}. \tag{5.1}$$

LEMMA 5.1. *The elementary projections $G^o_{\theta,b,a}$ for $\theta \in S^{n-1}$ fixed and $(b,a) \in K$ are mutually disjoint, i.e., $G^o_{\theta,b,a}G^o_{\theta,\beta,\alpha} = O$ whenever $(b,a) \neq (\beta,\alpha)$. In particular, if $K_0 \subseteq K$ is a finite subset, then*

$$G^o_{\theta,K_0} = \sum_{(b,a)\in K_0} G^o_{\theta,b,a}$$

is also an orthogonal projection.

Proof. Indeed, if we consider products of the projections $G^o_{\theta,b,a}$, then, for $f \in L^2(\mathbb{R}^n)$, we get

$$G^o_{\theta,b,a}G^o_{\theta,\beta,\alpha}f(y) = \int_{\mathbb{R}}\int_{\mathbb{R}} f(E_\theta y + s\theta)\overline{g_{\beta,\alpha}(s)}\ \overline{g_{b,a}(t)}g_{\beta,\alpha}(t)\,ds\,dt\ g_{b,a}(\langle y,\theta\rangle)$$

$$= \langle g_{\beta,\alpha}, g_{b,a}\rangle_{L^2(\mathbb{R})}P_g f(\theta, E_\theta y, \beta, \alpha)\ g_{b,a}(\langle y,\theta\rangle) = 0$$

whenever $(b,a) \neq (\beta,\alpha)$. The orthogonal projections are therefore mutually disjoint and the lemma follows immediately. □

We will show in Theorem 5.5 that functions in $L^2(\mathbb{R}^n)$ can be expanded in terms of elementary projections. The theorem will be proved using a number of elementary results which we will discuss first. We state the following lemma.

LEMMA 5.2. *Let $\{P_\nu\}_{\nu=1}^\infty$ be an increasing sequence of orthogonal projections on a Hilbert space H, i.e., $P_\mu P_\nu = P_\nu P_\mu = P_\mu$ for $\mu \leq \nu$. Then the following statements are equivalent:*

(1) *$P_\nu x \to x$ for all $x \in H$.*

(2) *$\bigcup_{\nu=1}^\infty \operatorname{ran} P_\nu \subseteq H$ dense.*

(3) *$\bigcap_{\nu=1}^\infty \ker P_\nu = (0)$.*

Next, we will construct an orthonormal basis in $L^2(\mathbb{R}^n)$ using products of elementary projections. If $\theta \perp \varphi$, then

$$G^o_{\theta,b,a} G^o_{\varphi,\beta,\alpha} f(y) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(E_\varphi E_\theta y + t\theta + s\varphi) \overline{g_{\beta,\alpha}(s)}\, \overline{g_{b,a}(t)}\, ds\, dt\, g_{\beta,\alpha}(\langle y, \varphi\rangle) g_{b,a}(\langle y, \theta\rangle).$$

Observe that $E_\theta E_\varphi = E_\varphi E_\theta$ is the orthoprojector onto the $(n-2)$-dimensional subspace $\{\theta, \varphi\}^\perp$. Moreover, it is immediate that the orthoprojectors $G^o_{\theta,b,a}$ and $G^o_{\varphi,\beta,\alpha}$ commute. The following two lemmas are direct consequences of these facts. The lemmas describe the so-called *separable wavelet analysis* which has been the first approach towards wavelet analysis of functions in several variables; see for example [Dau].

LEMMA 5.3. *Let $\theta_1, \dots, \theta_n \in S^{n-1}$ be mutually orthogonal unit vectors, and let $(b_j, a_j) \in \mathbb{H}$ for $j = 1, \dots, n$. Write*

$$\underline{\theta} = (\theta_1, \dots, \theta_n), \quad \underline{b,a} = ((b_1, a_1), \dots, (b_n, a_n)).$$

Next, define $F_{\underline{\theta,b,a}} = \prod_{j=1}^n g_{b_j,a_j}(\langle \cdot, \theta_j\rangle)$ and $G^o_{\underline{\theta,b,a}} = \prod_{j=1}^n G^o_{\theta_j,b_j,a_j}$. Then $F_{\underline{\theta,b,a}} \in L^2(\mathbb{R}^n)$, $\|F_{\underline{\theta,b,a}}\|_{L^2(\mathbb{R}^n)} = 1$, and $G^o_{\underline{\theta,b,a}} f = \langle f, F_{\underline{\theta,b,a}}\rangle_{L^2(\mathbb{R}^n)}\, F_{\underline{\theta,b,a}}$. In particular, the operator $G^o_{\underline{\theta,b,a}}$ is an orthogonal projector of rank one.

LEMMA 5.4. *Let $K \subseteq \mathbb{H}$ be a countable subset and let $g \in L^2(\mathbb{R})$ be a wavelet, such that $\{g_{b,a}\}_{(b,a)\in K}$ is an orthonormal basis of $L^2(\mathbb{R})$. If $\theta_1, \dots, \theta_n \in S^{n-1}$ are fixed mutually orthogonal unit vectors, then $\{F_{\underline{\theta,b,a}}\}_{\underline{b,a}\in K^n}$ is an orthonormal basis of $L^2(\mathbb{R}^n)$. In particular, if $f \in L^2(\mathbb{R}^n)$, then*

$$f = \sum_{\underline{b,a}\in K^n} \langle f, F_{\underline{\theta,b,a}}\rangle_{L^2(\mathbb{R}^n)}\, F_{\underline{\theta,b,a}}$$

and

$$\|f\|^2_{L^2(\mathbb{R}^n)} = \sum_{\underline{b,a}\in K^n} |\langle f, F_{\underline{\theta,b,a}}\rangle_{L^2(\mathbb{R}^n)}|^2.$$

We are now ready to prove the following result.

THEOREM 5.5. *If $f \in L^2(\mathbb{R}^n)$ and $\theta \in S^{n-1}$, then*

$$f = \sum_{(b,a)\in K} G^o_{\theta,b,a} f$$

converges in $L^2(\mathbb{R}^n)$. In particular,

$$\|f\|^2_{L^2(\mathbb{R}^n)} = \sum_{(b,a)\in K} \|G^o_{\theta,b,a} f\|^2_{L^2(\mathbb{R}^n)}.$$

Proof. By Lemma 5.2, it suffices to prove that $\bigcap_{(b,a)\in K} \ker\ G^o_{\theta,b,a} = (0)$. Indeed, let $(K_\nu)_{\nu=1}^\infty$ be an increasing sequence of finite subsets in K such that $\bigcup_{\nu=1}^\infty K_\nu = K$. If $P_\nu = \sum_{(b,a)\in K_\nu} G^o_{\theta,b,a}$, then $(P_\nu)_{\nu=1}^\infty$ is an increasing sequence of orthogonal projections on $L^2(\mathbb{R}^n)$. We now prove that $\bigcap_{\nu=1}^\infty \ker\ P_\nu = (0)$. Assume that $f \in \ker\ G^o_{\theta,b,a}$ for all $(b,a) \in K$. This means that for each $(b,a) \in K$,

$$G^o_{\theta,b,a} f(x + t\theta) = P_g f(\theta, x, b, a)\, g_{b,a}(t) = 0$$

for almost all $(x,t) \in \theta^\perp \times \mathbb{R}$. In particular, $P_g f(\theta, x, b, a) = 0$ for almost all $x \in \theta^\perp$. Let $\theta, \theta_2, \ldots, \theta_n$ be an orthonormal basis in $\mathbb{R}^n$ and write $\underline{\theta} = (\theta, \theta_2, \ldots, \theta_n)$. Then by Lemma 5.4,

$$f = \sum_{\underline{(b,a)} \in K^n} \langle f, F_{\underline{\theta,b,a}} \rangle_{L^2(\mathbb{R}^n)} \, F_{\underline{\theta,b,a}}.$$

On the other hand, for arbitrary $\underline{(b,a)} \in K^n$,

$$\langle f, F_{\underline{\theta,b,a}} \rangle_{L^2(\mathbb{R}^n)} = \int_{\mathbb{R}^n} f(\langle y, \theta\rangle \theta + \langle y, \theta_2\rangle \theta_2 + \cdots + \langle y, \theta_n \rangle \theta_n)$$

$$\times \, \overline{g_{b_1,a_1}(\langle y, \theta\rangle)} \, \overline{g_{b_2,a_2}(\langle y, \theta_2\rangle)} \cdots \overline{g_{b_n,a_n}(\langle y, \theta_n\rangle)} \, dy$$

$$= \int_{\theta^\perp} \int_{\mathbb{R}} f(x + t\theta) \overline{g_{b_1,a_1}(t)} \, dt \; \overline{g_{b_2,a_2}(\langle x, \theta_2\rangle)} \cdots \overline{g_{b_n,a_n}(\langle x, \theta_n\rangle)} \, dx.$$

The inner integral satisfies

$$\int_{\mathbb{R}} f(x + t\theta) \overline{g_{b_1,a_1}(t)} \, dt = P_g f(\theta, x, b_1, a_1) = 0$$

for almost all $x \in \theta^\perp$, and we arrive at $\langle f, F_{\underline{\theta,b,a}} \rangle_{L^2(\mathbb{R}^n)} = 0$. Since $\underline{(b,a)} \in K^n$ was arbitrary, we get $f = 0$. This proves the theorem. □

Let $(K_\nu)_{\nu=1}^\infty$ be an increasing sequence of finite sets such that $K_\nu \uparrow K$. We state the following corollary to Theorem 5.5.

COROLLARY 5.6. *Let $f \in L^2(\mathbb{R}^n)$; then $G^o_{\theta,K_\nu} f \to f$ in $L^2(\mathbb{R}^n)$ as $K_\nu \uparrow K$.*

Proof. Observe that, by Theorem 5.5,

$$\|G^o_{\theta,K_\nu} f - f\|^2_{L^2(\mathbb{R}^n)} = \sum_{(b,a) \in K/K_\nu} \|G^o_{\theta,b,a} f\|^2_{L^2(\mathbb{R}^n)} \to 0, \quad K_\nu \uparrow K. \qquad □$$

So far, we have fixed $\theta \in S^{n-1}$. Next, we will take averages over orthogonal projections corresponding to a finite number of distinct unit vectors. The idea is that averaging over approximations from several directions will improve the approximation result. Proposition 5.7 shows that this is the case in a certain sense to be explained below. Indeed, let $\Theta \subseteq S^{n-1}$ be a finite set of unit vectors and define for a fixed finite subset $K_0 \subseteq K$,

$$T^o_{\Theta,K_0} = \frac{1}{|\Theta|} \sum_{\theta \in \Theta} G^o_{\theta,K_0}, \tag{5.2}$$

where $|\Theta|$ denotes the number of elements in Θ. In order to understand the relevance of the following result, define an *approximation of the identity* as a sequence of operators $\{T_n\}_{n=1}^\infty$ on a Hilbert space H such that $T_n x \to x$ as $n \to \infty$ for all $x \in H$. Observe that for fixed n, the subspace ker T_n contains elements which are not approximated by T_n. The performance of the approximation of the identity $\{T_n\}_{n=1}^\infty$ is therefore measured by the null spaces of the operators in the sequence. The smaller these subspaces are, the better one may expect the performance to be. Proposition 5.7 states that the null space of the average T^o_{Θ,K_0} is contained in all of the null spaces of the orthoprojectors G^o_{θ,K_0}, $\theta \in \Theta$.

Proposition 5.7. *Let* $\Theta = \{\theta_1, \ldots, \theta_N\} \subseteq S^{n-1}$ *denote a finite set of directions,* $K_0 \subseteq K$ *finite, and define* T^o_{Θ,K_0} *as in* (5.2). *Then*

$$\ker\ T^o_{\Theta,K_0} = \bigcap_{j=1}^{N} \ker\ G^o_{\theta_j,K_0}.$$

The proposition is a consequence of the following rather general statement.

Lemma 5.8. *If* $P_1, \ldots, P_N$ *are orthogonal projections on a Hilbert space* H, *and if* $\alpha_1, \ldots, \alpha_N$ *are strictly positive numbers, and* $T = \sum_{j=1}^{N} \alpha_j P_j$, *then* $\ker\ T = \bigcap_{j=1}^{N} \ker\ P_j$.

Proof. If $x \in \bigcap_{j=1}^{N} \ker\ P_j$, then obviously $Tx = 0$. On the other hand, if $Tx = 0$, then

$$0 = \langle Tx, x\rangle_H = \sum_{j=1}^{N} \alpha_j \langle P_j x, x\rangle_H = \sum_{j=1}^{N} \alpha_j \|P_j x\|_H^2,$$

and hence $P_j x = 0$ for $j = 1, \ldots, N$. This proves the lemma. □

The self-adjoint operator T^o_{Θ,K_0} given by (5.2) need not be an orthogonal projection. Although we have described its kernel (and thereby its range), it is somewhat cumbersome to identify functions in $\ker\ T^o_{\Theta,K_0}$. It would help if we could identify the orthogonal projection onto this subspace. An approximation of this projector can be obtained using the Kacmarz method, as in [HS] for the X-ray transform.

6. Discrete wavelet X-ray transform: General case. We now study the general case in which $g \neq h$. The discretizations of the wavelet X-ray transform are now based on the existence of biorthogonal pairs of Riesz bases. Indeed, we will assume that g, h are a pair of wavelets for which there exists a countable subset $K \subseteq \mathbb{H}$ such that $(g_{b,a})_{(b,a)\in K}$ and $(h_{b,a})_{(b,a)\in K}$ are Riesz bases in $L^2(\mathbb{R})$ with Riesz bounds A_g, B_g and A_h, B_h, respectively. Moreover, we assume that the biorthogonality condition

$$\langle g_{b,a}, h_{\beta,\alpha}\rangle_{L^2(\mathbb{R})} = \begin{cases} 1, & (b,a) = (\beta,\alpha), \\ 0, & (b,a) \neq (\beta,\alpha), \end{cases}$$

holds. From the discussion on Riesz systems in the introduction, it follows that $A_g B_h = A_h B_g = 1$. Examples of biorthogonal pairs of Riesz bases are given in, for example, [Dau, CDF, CR]. These examples are all based on the countable set $K = \{(k2^j, 2^j) \mid k, j \in \mathbb{Z}\}$. In principle, orthonormal bases are to be preferred over biorthogonal systems of Riesz bases, as they allow for simpler implementation. On the other hand, the relaxed conditions on biorthogonal systems allow for a wider choice of wavelets, e.g., splines of higher order; see [Chu, CDF, CR].

In the general case when $g \neq h$, the elementary projections need not be orthogonal. The analogue to Theorem 5.5 in the general case reads as follows. Theorem 6.1 and Corollary 6.2 will be proved under the weaker condition that $(g_{b,a})_{(b,a)\in K}$ and $(h_{b,a})_{(b,a)\in K}$ are dual frames in $L^2(\mathbb{R})$. The frame bounds of $(g_{b,a})_{(b,a)\in K}$ are A_g, B_g; therefore the frame bounds of $(h_{b,a})_{(b,a)\in K}$ are given by $A_h = B_g^{-1}$ and $B_h = A_g^{-1}$; see the description of frames in the introduction. We recall the definition of an elementary projection from section 4:

$$G_{\theta,b,a} f = \int_{\mathbb{R}} f(E_\theta \cdot + t\theta)\overline{g_{b,a}(t)}\, dt\, h_{b,a}(\langle \cdot, \theta\rangle), \quad (b,a) \in K.$$

THEOREM 6.1. *Assume that $(g_{b,a})_{(b,a)\in K}$ and $(h_{b,a})_{(b,a)\in K}$ are dual frames in $L^2(\mathbb{R})$. If $f \in L^2(\mathbb{R}^n)$, then*

$$f = \sum_{(b,a)\in K} G_{\theta,b,a} f.$$

The series converges in $L^2(\mathbb{R}^n)$. Moreover, we find that

$$\frac{A_h}{\|h\|^2_{L^2(\mathbb{R})}} \sum_{(b,a)\in K} \|G_{\theta,b,a} f\|^2_{L^2(\mathbb{R}^n)} \leq \|f\|^2_{L^2(\mathbb{R}^n)} \leq \frac{B_h}{\|h\|^2_{L^2(\mathbb{R})}} \sum_{(b,a)\in K} \|G_{\theta,b,a} f\|^2_{L^2(\mathbb{R}^n)}.$$

Proof. If $f \in L^2(\mathbb{R}^n)$, observe that $f(x+t\theta) = \sum_{(b,a)\in K} \langle f(x+\cdot\theta), g_{b,a}\rangle_{L^2(\mathbb{R})} h_{b,a}(t)$ for almost all $x \in \theta^\perp$ and $t \in \mathbb{R}$. Since $(g_{b,a})_{(b,a)\in K}$ is a frame with Riesz bounds A_g, B_g, we get

$$A_g \|f(x+\cdot\theta)\|^2_{L^2(\mathbb{R})} \leq \sum_{(b,a)\in K} |\langle f(x+\cdot\theta), g_{b,a}\rangle_{L^2(\mathbb{R})}|^2 \leq B_g \|f(x+\cdot\theta)\|^2_{L^2(\mathbb{R})}.$$

Observe that

$$\|G_{\theta,b,a} f\|^2_{L^2(\mathbb{R}^n)} = \int_{\mathbb{R}^n} |P_g f(\theta, E_\theta y, b, a)|^2 |h_{b,a}(\langle y, \theta\rangle)|^2 \, dy$$

$$= \int_{\theta^\perp} |P_g f(\theta, x, b, a)|^2 \, dx \, \cdot \|h\|^2_{L^2(\mathbb{R})} = \int_{\theta^\perp} |\langle f(x+\cdot\theta), g_{b,a}\rangle_{L^2(\mathbb{R})}|^2 \, dx \, \cdot \|h\|^2_{L^2(\mathbb{R})}.$$

By Fubini's theorem, we get

$$\sum_{(b,a)\in K} \|G_{\theta,b,a} f\|^2_{L^2(\mathbb{R}^n)} = \int_{\theta^\perp} \sum_{(b,a)\in K} |\langle f(x+\cdot\theta), g_{b,a}\rangle_{L^2(\mathbb{R})}|^2 \, dx \, \cdot \|h\|^2_{L^2(\mathbb{R})}.$$

The theorem can now be proved using the equality

$$\|f\|^2_{L^2(\mathbb{R}^n)} = \int_{\theta^\perp} \|f(x+\cdot\theta)\|^2_{L^2(\mathbb{R})} \, dx. \qquad \square$$

We mention the following corollary without proof.

COROLLARY 6.2. *Let $(K_\nu)_{\nu=1}^\infty$ be an increasing sequence of finite sets in K such that $K_\nu \uparrow K$. If $f \in L^2(\mathbb{R}^n)$, then $G_{\theta,K_\nu} f \to f$ in $L^2(\mathbb{R}^n)$ as $K_\nu \uparrow K$.*

Next, we study null spaces of sums of elementary projections. From now on, we assume that $(g_{b,a})_{(b,a)\in K}$ and $(h_{b,a})_{(b,a)\in K}$ form a biorthogonal system of Riesz bases. We will use the results that were obtained for orthogonal elementary projections in the preceding section to deal with the general case. In order to do this, choose a wavelet ψ such that $(\psi_{b,a})_{(b,a)\in K}$ is an orthonormal basis in $L^2(\mathbb{R})$. Assume that $\|\psi\|_{L^2(\mathbb{R})} = 1$. We shall write

$$G^o_{\theta,b,a} f = P_\psi f(\theta, E_\theta \cdot, b, a) \psi_{b,a}(\langle \cdot, \theta\rangle).$$

Define the operator $S_{\theta,h}$ on $L^2(\mathbb{R}^n)$ by

$$S_{\theta,h} f = \sum_{(b,a)\in K} P_\psi f(\theta, E_\theta \cdot, b, a) h_{b,a}(\langle \cdot, \theta\rangle). \tag{6.1}$$

For each $\rho \in L^2(\theta^\perp)$, $S_{\theta,h}$ maps $f = \rho(E_\theta \cdot)\psi_{b,a}(\langle \cdot, \theta \rangle)$ to $S_{\theta,h} f = \rho(E_\theta \cdot) h_{b,a}(\langle \cdot, \theta \rangle)$. In this manner, $S_{\theta,h}$ is a lifted version of the boundedly invertible mapping on $L^2(\mathbb{R})$ which maps $(\psi_{b,a})_{(b,a)\in K}$ onto $(h_{b,a})_{(b,a)\in K}$. We now prove that $S_{\theta,h}$ itself is indeed a boundedly invertible operator on $L^2(\mathbb{R}^n)$.

LEMMA 6.3. *The operator $S_{\theta,h}$ defined in* (6.1) *is boundedly invertible on $L^2(\mathbb{R}^n)$. Its inverse is given by*

$$S_{\theta,h}^{-1} f = \sum_{(b,a)\in K} P_g f(\theta, E_\theta \cdot, b, a)\psi_{b,a}(\langle \cdot, \theta \rangle),$$

and the adjoint of $S_{\theta,h}$ reads

$$S_{\theta,h}^{*} f = \sum_{(b,a)\in K} P_h f(\theta, E_\theta \cdot, b, a)\psi_{b,a}(\langle \cdot, \theta \rangle).$$

Observe that $S_{\theta,h}^{-1} = S_{\theta,g}^{*}$. Moreover, $\|S_{\theta,h}\| = B_h^{1/2}$ and $\|S_{\theta,h}^{-1}\| = A_h^{-1/2}$.

Proof. We first prove that $S_{\theta,h}$ is bounded on $L^2(\mathbb{R}^n)$. Let $f \in L^2(\mathbb{R}^n)$; then

$$\|S_{\theta,h} f\|_{L^2(\mathbb{R}^n)}^2 = \left\| \sum_{(b,a)\in K} P_\psi f(\theta, E_\theta \cdot, b, a) h_{b,a}(\langle \cdot, \theta, \rangle) \right\|_{L^2(\mathbb{R}^n)}^2$$

$$\leq B_h \sum_{(b,a)\in K} \|P_\psi f(\theta, E_\theta \cdot, b, a)\|_{L^2(\theta^\perp)}^2 = B_h \sum_{(b,a)\in K} \|G_{\theta,b,a}^{o} f\|_{L^2(\mathbb{R}^n)}^2 = B_h \|f\|_{L^2(\mathbb{R}^n)}^2.$$

In order to prove that the norm of $S_{\theta,h}$ equals $B_h^{1/2}$, let $\varepsilon > 0$ and let $(c_{(b,a)})_{(b,a)\in K} \in \ell^2(K)$ such that

$$\left\| \sum_{(b,a)\in K} c_{(b,a)} h_{b,a} \right\|_{L^2(\mathbb{R})}^2 > (B_h - \varepsilon) \sum |c_{(b,a)}|^2.$$

Let $\rho \in L^2(\theta^\perp)$ with unit norm, and let $f = \sum_{(b,a)\in K} c_{(b,a)} \rho(E_\theta \cdot)\psi_{b,a}(\langle \cdot, \theta \rangle)$. It is now easy to verify that

$$\frac{\|S_{\theta,h} f\|_{L^2(\mathbb{R}^n)}^2}{\|f\|_{L^2(\mathbb{R}^n)}^2} \geq B_h - \varepsilon.$$

In the same fashion, it can be shown that $\|S_{\theta,g}^{*}\| = A_h^{-1/2}$. Using Theorem 6.1, one proves that $S_{\theta,g}^{*} S_{\theta,h} = S_{\theta,h} S_{\theta,g}^{*} = I$. □

We state the following lemma without proof.

LEMMA 6.4.

$$S_{\theta,h}^{-1} G_{\theta,b,a} S_{\theta,h} = G_{\theta,b,a}^{o}.$$

The following results are obtained easily now.

LEMMA 6.5.

$$\ker G_{\theta,b,a} = S_{\theta,h} \ker G_{\theta,b,a}^{o}.$$

Corollary 6.6. *Define* $G_{\theta,K_0} = \sum_{(b,a)\in K_0} G_{\theta,b,a}$*; then*

$$\ker G_{\theta,K_0} = \bigcap_{(b,a)\in K_0} \ker G_{\theta,b,a}.$$

As in the previous section, we shall study the approximation of functions by averaging over approximations from several directions. In analogy to (5.2), we define the operator

$$T_{\Theta,K_0} = \frac{1}{|\Theta|} \sum_{\theta\in\Theta} G^*_{\theta,K_0} G_{\theta,K_0}, \tag{6.2}$$

where both $\Theta \subset S^{n-1}$ and $K_0 \subset K$ are finite subsets. Observe that if G_{θ,K_0} is an orthogonal projection, then $G^*_{\theta,K_0} G_{\theta,K_0} = G_{\theta,K_0}$; therefore the definition of T_{θ,K_0} here is consistent with the one given in (5.2). The proof of the following proposition is achieved along the same lines as the proof of Proposition 5.7.

Proposition 6.7. *Let* $\Theta = \{\theta_1, \ldots, \theta_N\} \subseteq S^{n-1}$ *denote a finite set of directions,* $K_0 \subseteq K$ *finite, and define* T_{Θ,K_0} *as in* (6.2)*. Then*

$$\ker\ T_{\Theta,K_0} = \bigcap_{j=1}^{N} \ker\ G_{\theta_j,K_0}.$$

It is not known to the author whether a result as in the Proposition 6.7 holds true for an operator of the form

$$\widetilde{T}_{\Theta,K_0} = \frac{1}{|\Theta|} \sum_{\theta\in\Theta} G_{\theta,K_0}.$$

Observe that Lemma 5.8 does not hold for sums of not necessarily orthogonal projections.

REFERENCES

[AM] J.P. Antoine and R. Murenzi, *Two-Dimensional Directional Wavelets and the Scale-Angle Representation*, Tech. Report UCL-IPT-95-03, Department of Theoretical and Mathematical Physics, Université Catholique de Louvain, Louvain, Belgium, 1995.

[BW] C.A. Berenstein and D.F. Walnut, *Local inversion Radon transform in even dimensions using wavelets*, in 75 Years of Radon Transform, S. Gindikin and P. Michor, eds., International Press, Cambridge, MA, 1994, pp. 45–69.

[Chu] C.K. Chui, *An Introduction to Wavelets*, Academic Press, Boston, MA, 1992.

[CDF] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[CR] A. Cohen and R.D. Ryan, *Wavelets and Multiscale Signal Processing*, Chapman & Hall, London, 1995.

[Dau] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Math. 61, SIAM, Philadelphia, PA, 1992.

[Dau2] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[Dea] S.R. Deans, *The Radon Transform and Some of Its Applications*, John Wiley, New York, 1983.

[DL] K.A. Dines and R.J. Lytle, *Computerized geophysical tomography*, Proc. IEEE, 67 (1979), pp. 1065–1073.

[FKV] L. Faqi, M.M.N. Kabir, and D.J. Verschuur, *Seismic processing using the wavelet and the Radon transform*, J. Seismic Exploration, 4 (1995), pp. 375–390.

[FKV2] L. Faqi, M.M.N. Kabir, and D.J. Verschuur, *Cascaded application of the linear Radon and wavelet transform in preprocessing*, in Proc. Annual Meeting Society Exploration Geophysicists, Expanded Abstracts, Houston, TX, 1995, pp. 1373–1376.

[HS] C. Hamaker and D.C. Solmon, *The angles between null spaces of X rays*, J. Math. Anal. Appl., 62 (1978), pp. 1–23.

[Hol] M. Holschneider, *Wavelets: An Analysis Tool*, Clarendon Press, Oxford, 1995.

[KS] G. Kaiser and R.F. Streater, *Windowed Radon transforms, analytic signals, and the wave equation*, in Wavelets: A Tutorial in Theory and Applications, C.K. Chui, ed., Academic Press, New York, 1992, pp. 399–441.

[Koo] T.H. Koornwinder, *The continuous wavelet transform*, in Wavelets: An Elementary Treatment of Theory and Applications, T.H. Koornwinder, ed., World Scientific, River Edge, NJ, 1993, pp. 27–48.

[Mey] Y. Meyer and R.D. Ryan, *Wavelets: Algorithms and Applications*, SIAM, Philadelphia, PA, 1993.

[Nat] F. Natterer, *The Mathematics of Computerized Tomography*, John Wiley, Chichester, NY, 1986.

[OD] T. Olson and J. DeStefano, *Wavelet localization of the Radon Transform*, IEEE Trans. Signal Process., 42 (1994), pp. 2055–2067.

[RV] O. Rioul and M. Vetterli, *Wavelets and signal processing*, IEEE Signal Process. Magazine, 8 (1991), pp. 14–38.

[Rob] E.A. Robinson, *Spectral approach to geophysical inversion by Lorentz, Fourier, and Radon transforms*, Proc. IEEE, 70 (1984), pp. 1039–1054.

[Ru] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1986.

[SSW] K.T. Smith, D.C. Solmon, and S.L. Wagner, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. Amer. Math. Soc. (N.S.), 82 (1977), pp. 1227–1270.

[Sol] D.C. Solmon, *The X-ray transform*, J. Math. Anal. Appl., 56 (1976), pp. 61–83.

[Tak] T. Takiguchi, *On invertibility of the windowed Radon transform*, J. Math. Sci. Univ. Tokyo, 2 (1995), pp. 621–636.

[You] R.M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[Wal] J.S. Walker, *Fourier analysis and wavelet analysis*, Notices Amer. Math. Soc., 44 (1997), pp. 658–670.

[WD] A.L. Warrick and P.A. Delaney, *A wavelet localized Radon transform*, in SPIE Proceedings 2569, Wavelet Applications in Signal and Image Processing III, San Diego, CA, 1995, pp. 632–643.

[Zui] R.A. Zuidwijk, *The Wavelet X-Ray Transform*, CWI-report PNA-R9703, Center for Mathematics and Computer Science, Amsterdam, 1997.

[Zui2] R.A. Zuidwijk, *The discrete and continuous wavelet X-ray transform*, in SPIE Proceedings 3169, Wavelet Applications in Signal and Image Processing V, San Diego, CA, 1997, pp. 357–366.

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 431–454

A GEOMETRIC APPROACH TO SINGULARLY PERTURBED NONLOCAL REACTION-DIFFUSION EQUATIONS*

AMITABHA BOSE†

Abstract. In the context of a microwave heating problem, a geometric method to construct a spatially localized, 1-pulse steady-state solution of a singularly perturbed, nonlocal reaction-diffusion equation is introduced. The 1-pulse is shown to lie in the transverse intersection of relevant invariant manifolds. The transverse intersection encodes a consistency condition that all solutions of nonlocal equations must satisfy. An oscillation theorem for eigenfunctions of nonlocal operators is established. The theorem is used to prove that the linear operator associated with the 1-pulse solution possesses an exponentially small principal eigenvalue. The existence and instability of n-pulse solutions is also proved. A further application of the theory to the Gierer–Meinhardt equations is provided.

Key words. nonlocal reaction-diffusion equation, geometric singular perturbation theory, transversality, metastability, eigenvalues

AMS subject classifications. 35K57, 35K60, 34C10

PII. S0036141098342556

1. Introduction.

This paper is concerned with establishing a geometric method to analyze singularly perturbed, nonlocal reaction-diffusion equations. Such equations arise in microwave heating applications [18], activator-inhibitor chemical systems [20], thermistor [19], and ballast resistor problems [3], among other places. Nonlocal equations also are of interest because a higher-dimensional system can often be recast into a lower-dimensional nonlocal system [5, 7, 9, 14]. Existence of stationary solutions for scalar nonlocal equations has been considered by [3, 7, 9, 18, 19]. Stability of solutions for scalar nonlocal equations has been studied by [1, 3, 5, 10, 11, 12, 14].

Various methods for showing existence of solutions to nonlocal, boundary value problems have been employed. In [3, 9], it is shown that a homogeneous steady-state solution becomes unstable and a certain bifurcation occurs, thereby yielding a new steady-state or time-periodic solution. Lacey uses Picard iteration to show that nonhomogeneous steady-state solutions exist for equations modeling thermistors [19]. These methods establish the existence of solutions, but are not constructive in the sense that they give little information about the structure of solutions and where they lie in an appropriate phase space. Alternatively, asymptotic methods can be used to formally construct solutions [18]. However, asymptotic analysis does not actually yield a proof that a solution exists, and without a proof, one cannot be certain that what is constructed by formal asymptotics actually corresponds to the asymptotics of the true solution. Indeed, there are examples in the literature in which what appear to be asymptotic approximations of solutions are derived, but in fact there are no true solutions nearby at all. In the spirit of the approach that we shall take, Doelman and Rottschäfer [7] have studied a nonlocal reduction of the Ginzburg–Landau equations from a geometric point of view. The main purpose of their work is to compare a nonlocal model to a singularly perturbed one to determine whether the former is a

*Received by the editors July 28, 1998; accepted for publication (in revised form) May 17, 1999; published electronically January 10, 2000. This work was supported in part by the New Jersey Institute of Technology under grant 421540.
http://www.siam.org/journals/sima/31-2/34255.html

†Department of Mathematical Sciences, Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark, NJ 07102 (bose@m.njit.edu).

good approximant of the latter. As a result, their nonlocal equations do not involve singular perturbations.

In this paper, in the context of a microwave heating problem, it is shown how to use geometric singular perturbation theory to construct spatially localized 1-pulse solutions for a nonlocal boundary value problem. A 1-pulse solution contains one spatially localized local maxima. This solution is shown to lie in the transverse intersection of relevant invariant manifolds. Two manifolds intersect transversely, if at a point of intersection, the tangent spaces of these manifolds span the ambient space. Solutions to nonlocal equations must satisfy a so-called consistency condition, which we discuss in detail below. An important byproduct of this work is the development of a geometric method to determine which trajectories in phase space satisfy the consistency condition. We show how to replace the scalar nonlocal boundary value problem with a higher-dimensional local boundary value problem in which the consistency constraint has been embedded.

Transversality is obtained as a direct result of the higher dimensionality. In fact, our analysis shows that transversality implies satisfaction of the consistency condition. Furthermore, since transversality of manifolds implies local uniqueness of intersections, we also obtain local uniqueness of the 1-pulse solution. This means that in a neighborhood of phase space of a 1-pulse solution, there are no other steady-state solutions. We also prove the existence of spatially localized n-pulse solutions.

Freitas [10, 11, 12] has obtained extensive results on the stability of scalar nonlocal reaction-diffusion equations. He shows how to locate the spectrum of a nonlocal linear operator by seeing how the spectrum of a related local operator changes under perturbations. Here, we locate the spectrum of a nonlocal version of a standard Sturm–Liouville operator. We show that the 1-pulse is metastable in that the nonlocal operator possesses an exponentially small principal eigenvalue. For $n \geq 2$, the n-pulses are unstable. The primary tool we employ is an oscillation theorem found in Bose and Kriegsmann [1]. Using Freitas' results, we show that this theorem holds under more general circumstances than those found in [1].

The equation of interest arises in microwave heating applications. A spatially localized hot spot forms in a thin ceramic fiber when it is microwave heated in a highly resonant, single mode cavity. The spot forms along the axis of the sample and begins to propagate [21]. In most cases the spot eventually becomes stationary, thereby leaving a localized region of the fiber at a dramatically higher temperature than the rest. Kriegsmann [18] derives the following nonlocal reaction-diffusion equation to model this phenomena:

$$U_t = \epsilon^2 U_{xx} + \frac{pf(U)}{1 + c^2(\int_0^1 f(U)\ dx)^2} - h(U), \tag{1.1}$$

$$U_x(0,t) = U_x(1,t) = 0.$$

Here $0 \leq x \leq 1$, U corresponds to the dimensionless temperature along the fiber and is assumed to be nonnegative; p is the dimensionless power which is proportional to the square of the amplitude of the mode which excites the cavity; and c lumps several physical parameters together. The function $h(U)$ models heat loss at the surface of the fiber due to convection and radiation and satisfies $h(0) = 0$, $h'(U) > 0$. The function $f(U)$ represents the effective electrical conductivity of a low-loss ceramic, such as alumina, and satisfies $f(0) = 1$, $f'(U) > 0$ and $f(U)$ grows faster than $h(U)$

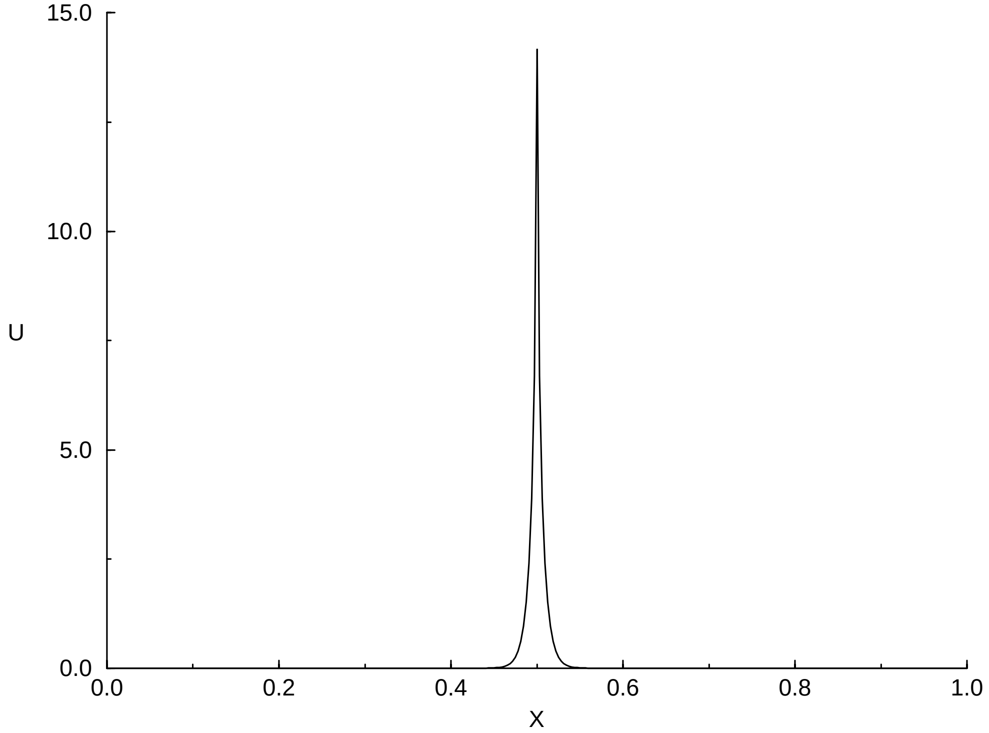


FIG. 1.1. *Kriegsmann's* 1*-pulse solution with* $\epsilon = 0.01$*,* $p = 1.0$*,* $c = 0.01$*,* $\beta = 0.01$*,* $c_1 = 1.0$*.*

for sufficiently large values of U. Both f and h are assumed to be sufficiently smooth. The nonlocal term models the detuning effect the heated fiber has upon the cavity. The diffusion constant ϵ is the aspect ratio of the fiber and in practice is much less than one, thereby making (1.1) singularly perturbed.

In [18], Kriegsmann chooses $f(U) = e^{c_1 U}$, $c_1 > 0$ and $h(U) = 2(U+\beta[(U+1)^4-1])$, $\beta \ll 1$. There, he formally constructs a localized, 1-pulse, steady-state solution ($U_t = 0$) using matched asymptotic expansions (see Figure 1.1). Kriegsmann shows that for ϵ sufficiently small, the 1-pulse possesses the following characteristics:

(C1) The solution is symmetric about $x = 1/2$.

(C2) It is nearly constant and close to zero valued on most of $[0, 1]$.

(C3) On a small interior layer centered around $x = 1/2$, the solution attains a maximum value that tends to ∞ as $\epsilon \to 0$. The value of the nonlocal integral term also tends to ∞ in this limit.

Recently, Bose and Kriegsmann [1] proved that this solution is metastable. The solution is, in fact, unstable, but perturbations of the solution decay to the solution or a translate of it, which then persists for exponentially long amounts of time.

In this paper, we study (1.1) with $f(U) = 1 + U^2$ and $h(U) = 2U$. Our choice of these functions is motivated by the fact that they are the simplest functions for which solutions of (1.1) retain the qualitative features of those found in [18]. Moreover, this choice simplifies the analysis so as to focus on the geometric approach. See Remark 3.6 for a discussion on the effect of including $O(\beta)$ terms in the function $h(U)$. To construct a steady-state 1-pulse solution of (1.1), we study the following boundary value problem:

$$\epsilon^2 U_{xx} + \frac{p(1+U^2)}{1+c^2(\int_0^1 1+U^2\ dx)^2} - 2U = 0, \tag{1.2}$$

$$U_x(0) = U_x(1) = 0.$$

Let $I = \int_0^1 U^2\ dx$. Replacing $\int_0^1 1+U^2\ dx$ by $1+I$, we note that the value I is determined by the solution itself. Thus a solution of (1.2) must satisfy the *consistency condition*

$$I_\star = \int_0^1 U^2(x, I_\star)\ dx. \tag{1.3}$$

The prescribed value $I_\star$ must be the value that is determined by computing the integral of U^2 along a trajectory of (1.2). This motivates the introduction of an auxiliary variable $V(x) = \int_0^x U^2\ dx$.

We recast (1.2) as the following system of first-order equations, where $' = d/dx$:

$$\begin{aligned} \epsilon U' &= W, & V' &= U^2, \\ \epsilon W' &= 2U - \tfrac{p(1+U^2)}{1+c^2(1+I)^2}, & I' &= 0. \end{aligned} \tag{1.4}$$

Note that $V(1) = \int_0^1 U^2\ dx$. This formulation, in a very natural way, recasts (1.2)–(1.3) into the boundary value problem (1.4), subject to the boundary conditions

$$\begin{aligned} (U, W, V, I) &= (U(0), 0, 0, I_\star) \text{ at } x = 0, \\ (U, W, V, I) &= (U(1), 0, I_\star, I_\star) \text{ at } x = 1. \end{aligned} \tag{1.5}$$

The new consistency condition is $V(1) = I_\star$. The analysis will show that $I_\star$ is unique, as a priori, this is not obvious. The values of $U(0)$ and $U(1)$ will also need to be determined. Since $\int_0^1 1+U^2\ dx = 1+I$ is simply a number, by phase plane considerations, any pulse solution of (1.4)–(1.5) will necessarily be symmetric about $x = 1/2$. Thus $U(0) = U(1)$. Also, the trivial equation $I' = 0$ in (1.4) is necessary, as the unique value $I_\star$ will be determined by transversality with respect to I.

We employ geometric singular perturbation theory to construct a 1-pulse solution of (1.4)–(1.5) [8, 15]. This theory involves finding solutions to sets of reduced equations obtained by formally setting $\epsilon = 0$ in appropriately scaled versions of (1.4). These solutions are then pieced together to form a singular solution. If the singular solution lies in certain manifolds which satisfy relevant transversality conditions, then an actual solution for $\epsilon \ll 1$ is obtained. Transverse intersections persist under perturbation which allows the $\epsilon = 0$ results to be extended to ϵ small. The first theorem that we prove is the following.

THEOREM 1.1. (a) *For ϵ sufficiently small, there exists a locally unique, symmetric* 1*-pulse solution U_1 of* (1.4)–(1.5)*. The maximum value U_{max} of this solution and the unique value $I_\star$ of the nonlocal term are given by*

$$U_{max} = 3c^2 I_\star^2/p + O(\epsilon), \qquad I_\star = \left(\frac{p^2}{12\sqrt{2}c^4\epsilon}\right)^{1/3} + O(\epsilon).$$

(b) *Let ϵ_1 be sufficiently small such that a symmetric* 1*-pulse solution exists as in* (a)*. Then for $\epsilon = \epsilon_1/n$, there exists a symmetric n-pulse solution U_n of* (1.4)–(1.5).

Stability of these solutions with respect to arbitrary perturbations is an important property for any physically realizable solution. For nonlocal equations of the type under consideration, Chafee [3] has shown that linear stability implies asymptotic stability. If a solution is asymptotically stable, then arbitrary perturbations of the solution decay in an appropriate function space. The main result to be proved here is that the 1-pulse solution is metastable, as is the case for the 1-pulse constructed in [18] and analyzed in [1].

THEOREM 1.2. (a) *The* 1*-pulse solution* U_1 *is a metastable solution of* (1.1). (b) *For* $n \geq 2$*,* n*-pulse solutions* U_n *are unstable solutions of* (1.1) *with principal eigenvalue bounded away from the origin as* $\epsilon \to 0$*.*

We now give an outline of the paper. Due to criterion (C3) of Kriegsmann's solution, it turns out that (1.4) is not the correctly scaled version of the equations with which to work. In section 2, we rescale (1.4) to obtain the correct set of equations, together with relevant sets of reduced equations. In section 3, we construct a singular 1-pulse solution and show that it perturbs to yield an actual 1-pulse solution for ϵ sufficiently small. We also obtain n-pulse solutions using a simple rescaling argument. In section 4, we prove Theorem 1.2 concerning the stability of solutions. Section 5 contains numerical simulations of the full time-dependent equations. The numerically obtained values for U_{max} and $I_\star$ are shown to agree closely with the theoretically predicted ones. In section 6, we give a second application of our theory to a special limit of the Gierer–Meinhardt equations [13] which describe biological pattern formation. In this limit, the system of two reaction-diffusion equations can be reduced to a scalar nonlocal reaction-diffusion equation of the type considered above. A brief discussion concludes the paper.

2. The singular solution.

2.1. Scalings. The asymptotic analysis of Kriegsmann [18] shows that the maximum value of the 1-pulse solution of (1.1) tends to infinity as ϵ tends to 0. Moreover, the value of the nonlocal term I also tends to infinity in this limit. Numerical simulations of (1.1), with either set of nonlinearities discussed above, show that the main contribution to the nonlocal term occurs on the small interior layer around $x = 1/2$. A set of outer and inner equations associated with (1.4) can naively be derived in an attempt to capture this behavior. An outer set of equations is found simply by setting $\epsilon = 0$ in (1.4). An inner set of equations is obtained by rescaling the spatial variable x in a neighborhood of $1/2$ by using $\xi = (x - 1/2)/\epsilon$ and then setting $\epsilon = 0$ in the ensuing equations. The problem with this scaling is that $dV/d\xi = 0$. Thus there would be no contribution to the nonlocal integral term over the inner solution.

Following ideas similar to those found in [6], we rescale both the spatial and dependent variables in (1.4). Let $u = \epsilon^a U$, $w = \epsilon^a W$, $v = \epsilon^b V$, $Z = \epsilon^b I$, and $\xi = (x - 1/2)/\epsilon$. We use a capital Z for the new scaling of I to emphasize that transversality occurs with respect to this parameter. It is clear why U and W need to be scaled by the same power of ϵ. That V and I also need to be scaled by equal powers of ϵ follows from the fact that the consistency condition requires $V(1) = I$. Introducing these scalings in (1.4) with $\dot{} = d/d\xi$ yields

$$
\begin{aligned}
&\dot{u} = w, && \dot{v} = \epsilon^{b+1-2a} u^2, \\
&\dot{w} = 2u - p\, \tfrac{\epsilon^{a+2b} + \epsilon^{2b-a} u^2}{\epsilon^{2b} + c^2(\epsilon^b + Z)^2}\ , && \dot{Z} = 0, \\
&\dot{Z} = 0. &&
\end{aligned}
\tag{2.1}
$$

We require that all nonconstant terms in (2.1) be $O(1)$. We do this in order to make sure that both the linear (u) and the nonlinear (u^2) terms in the w component of the vector field are $O(1)$ so that the fast subsystem possesses a homoclinic orbit. Doing so implies $2b = a$ and $2a = b + 1$. Thus $a = 2/3$ and $b = 1/3$. Next, scale back to the x-variable, introduce $y = x$, and append the equation $dx/dy = 1$ to allow the invariant manifolds, defined below, to explicitly contain a spatial component. The set of equations then becomes

$$\begin{aligned} \epsilon du/dy &= w, & dZ/dy &= 0, \\ \epsilon dw/dy &= 2u - p\tfrac{\epsilon^{4/3}+u^2}{\epsilon^{2/3}+c^2(\epsilon^{1/3}+Z)^2}, & dx/dy &= 1, \\ \epsilon dv/dy &= u^2. \end{aligned} \tag{2.2}$$

In these new scalings, the consistency condition (1.3) becomes

$$\epsilon Z_\star = \int_0^1 u^2(y, Z_\star)\ dy. \tag{2.3}$$

The boundary conditions (1.5) transform to

$$\begin{aligned} (u, w, v, Z, x) &= (u(0), 0, 0, Z_\star, 0), \\ (u, w, v, Z, x) &= (u(1), 0, Z_\star, Z_\star, 1). \end{aligned} \tag{2.4}$$

In terms of these boundary conditions, the consistency condition is also recognized as $v(1) = Z_\star$. As before the symmetry of the 1-pulse implies $u(0) = u(1)$, and these values along with $Z_\star$ will need to be determined by the analysis.

2.2. Solutions to reduced equations and singular solutions. The inner and outer sets of equations associated with (2.2) are now easy to obtain. To derive the outer equations, set $\epsilon = 0$ in (2.2):

$$\begin{aligned} 0 &= w, & dZ/dy &= 0, \\ 0 &= 2u - p\tfrac{u^2}{c^2Z^2}, & dx/dy &= 1, \\ 0 &= u^2. \end{aligned} \tag{2.5}$$

To obtain the inner inequations which incorporate the symmetry of the 1-pulse, condition (C1), we rescale (2.2) in a neighborhood of $y = 1/2$ using $\xi = (y - 1/2)/\epsilon$, and set $\epsilon = 0$:

$$\begin{aligned} \dot{u} &= w, & \dot{Z} &= 0, \\ \dot{w} &= 2u - p\ \tfrac{u^2}{c^2Z^2}\ , & \dot{x} &= 0, \\ \dot{v} &= u^2. \end{aligned} \tag{2.6}$$

Solutions to (2.5) are easily found as both u and w are forced to be zero by the vector field. The values of v and Z, however, are unspecified, but constant for the outer equations. Solutions to the outer equations capture the behavior described in condition (C2). Solutions to (2.6) are also easily obtained. Note that the u and w equations are decoupled from the v equation and that Z acts as a parameter in these equations. For each value of Z, the $u - w$ equations are Hamiltonian with critical points $(0, 0)$ and $(2c^2Z^2/p, 0)$. The origin is a saddle point and the other critical point is a center. The $u - w$ phase plane is shown in Figure 2.1. The value of v along an

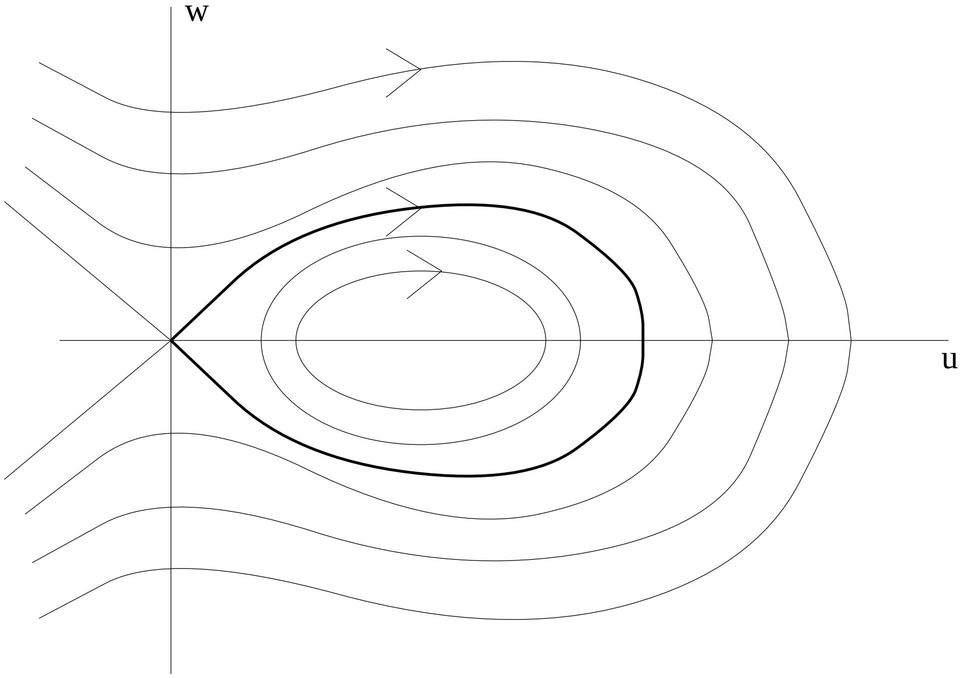


FIG. 2.1. *The phase plane for the $u - w$ equations of* (2.6). *The homoclinic solution is the darker curve.*

inner solution is determined by integrating u^2 along a trajectory in the $u - w$ phase plane. The homoclinic solution of the inner equations (2.6) captures the behavior described in (C3).

The singular solution can now be constructed. By definition, $v(0) = 0$. Thus at $y = 0$, only Z is unspecified. Since Z can vary, there exists a one-parameter family of singular solutions. We describe one particular member of this family. Fix $Z > 0$. The first piece of the singular trajectory is a solution of (2.5) that connects $(0, 0, 0, Z, 0)$ to $(0, 0, 0, Z, 1/2)$ in (u, w, v, Z, x) space. The second piece is the solution of (2.6) that connects $(0, 0, 0, Z, 1/2)$ at $\xi = -\infty$ to $(0, 0, v, Z, 1/2)$ at $\xi = \infty$. In the u and w components, this singular piece corresponds to the homoclinic orbit pictured in Figure 2.1. Since the value of v at $\xi = -\infty$ is different than at $\xi = +\infty$, the inner piece is actually a heteroclinic orbit in the full five-dimensional phase space. The third and final piece of the singular trajectory is a solution of (2.5) from $(0, 0, v, Z, 1/2)$ to $(0, 0, v, Z, 1)$.

With these scalings, the consistency condition (2.3) reduces to a particularly simple form. In the outer equations (2.5), $u = 0$. Thus, the outer solutions contribute nothing to the integral. The $\epsilon = 0$ value of $Z_\star$, denoted by Z_0, is determined solely by the inner equations and is given by

$$Z_0 = \int_{-\infty}^{\infty} u^2(\xi, Z_0)\, d\xi. \tag{2.7}$$

3. Invariant manifolds and transversality. We now pick out a unique singular solution which satisfies (2.7) from the one-parameter family of singular 1-pulse solutions. We then extend the analysis to ϵ small. To do both, we recast the above analysis into the language of invariant manifolds. Our analysis relies on the seminal work of Fenichel [8] on the persistence of invariant manifolds. See Jones [15] and the references therein for a thorough exposition of the theory and some of its applications.

Following Tin, Kopell, and Jones [22], we define manifolds of points which, respectively, satisfy the boundary conditions (2.4) at $y = 0$ and 1. We then flow the $y = 0$ boundary manifold forward to determine whether it intersects the boundary manifold at $y = 1$. Denote the flows of (2.5) and (2.6) as the outer and inner flows, respectively. These flows are used to track the evolution of the $y = 0$ boundary manifold over different pieces of the singular solution.

The $y = 0$ boundary manifold is defined by

$$B_0 = \{(u, w, v, Z, x) : u = w = v = x = 0\}.$$

Thus B_0 is a one-dimensional curve consisting solely of different Z values along which u, w, v, and x are restricted to 0. At $y = 1$, we define two different boundary manifolds. First, let

$$B_R = \{(u, w, v, Z, x) : u = w = 0, x = 1\}.$$

The manifold B_R is two-dimensional as both v and Z are free and by definition positive. It contains no information about the consistency condition. Enforcing $v(1) = Z$ restricts B_R to the following one-dimensional submanifold:

$$B_1 = \{(u, w, v, Z, x) : u = w = 0, v = Z, x = 1\}.$$

At $y = 1/2$, the jump off and touch down curves are defined. These are curves along which the outer and inner flows must match. The jump off curve is

$$J_0 = \{(u, w, v, Z, x) : u = w = v = 0, x = 1/2\}$$

and the touch down curve is

$$T_0 = \left\{(u, w, v, Z, x) : u = w = 0, v = \int_{-\infty}^{\infty} u^2(\xi, Z)\, d\xi, x = 1/2\right\}.$$

The touch down curve is determined by flowing J_0 forward under the inner flow. In particular, it contains no information about the consistency condition. In the next subsection, we will define an analogous curve T_c which will encode the consistency condition.

Denote by $B_0 \cdot y$ the two-dimensional manifold formed by flowing B_0 forward. Note that $J_0 = B_0 \cdot y|_{y=1/2^-}$, so the outer flow transversely intersects the jump off curve on the slow manifold. This is essential for the perturbation result later for ϵ small. Now use the inner flow to follow $J_0 = B_0 \cdot 1/2^-$. For each point on J_0, there exists a homoclinic solution of the inner $u - w$ equations (2.6). As mentioned earlier, considering the full set of inner equations in a five-dimensional phase space, this is actually a heteroclinic solution in that the values of v and Z on the jump off curve J_0 and touch down curve T_0 are dramatically different; see Figure 3.1. Thus the inner flow defines a two-dimensional (sheet) manifold of heteroclinic orbits, which connects J_0 to T_0. Therefore T_0 is also one-dimensional. Lastly, flow T_0 forward to $y = 1$ using the outer flow to obtain $T_0 \cdot 1$. The curve $T_0 \cdot 1$ is the image of B_0 flowed under the appropriate outer and inner equations. Thus we define $B_0 \cdot 1 = T_0 \cdot 1$. By

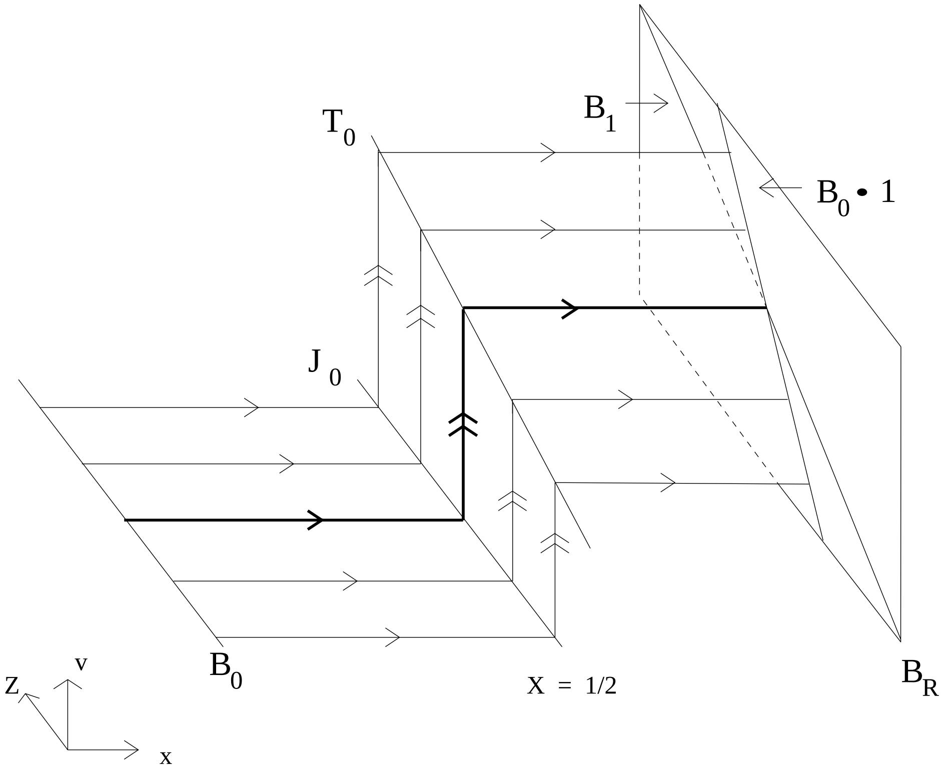


FIG. 3.1. *A graphical representation of the singular manifolds, curves, and flows, projected into (v, Z, x) space. The darker curve depicts the unique singular solution that satisfies the consistency condition.*

construction, $B_0 \cdot 1 \cap B_R \neq \emptyset$. If there is a singular 1-pulse solution, then it must satisfy the following geometric version of (2.7):

$$B_0 \cdot 1 \cap B_1 \neq \emptyset. \tag{3.1}$$

Further, if $B_0 \cdot 1$ transversely intersects B_1 in $\mathbf{R^2}$, then there exists a unique singular 1-pulse solution which satisfies (2.7). The following lemma establishes the transversality.

LEMMA 3.1. *The curve T_0 transversely intersects the line $v = Z$ at a unique point in (Z, v) space.*

Proof. It is easy to verify that

$$u(\xi) = \frac{3c^2 Z^2}{p} \operatorname{sech}^2 \frac{\xi}{\sqrt{2}}, \quad w(\xi) = \dot{u} \tag{3.2}$$

solves the $u - w$ equations of (2.6). Let $v(Z)$ be defined by the first equation of (3.3), below. It is easily checked that

$$v(Z) = \int_{-\infty}^{+\infty} u^2(\xi, Z)\, d\xi = \tfrac{12\sqrt{2}c^4 Z^4}{p^2}. \tag{3.3}$$

The graph of (3.3) is exactly T_0 projected onto (Z, v) space; see Figure 3.2. Note that it intersects the line $v = Z$ at exactly one point. This intersection value is calculated

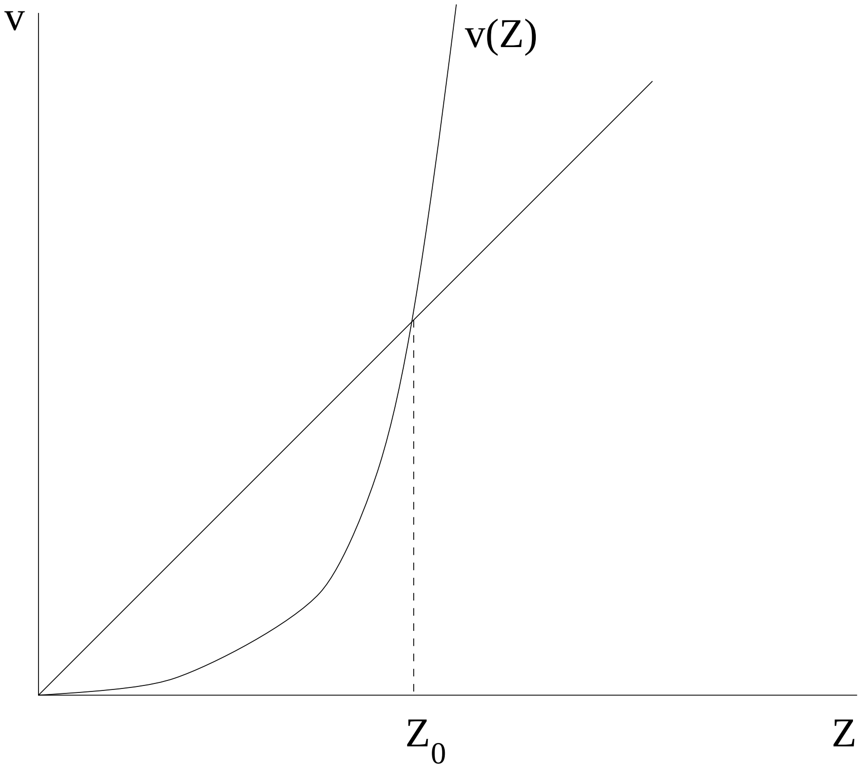


FIG. 3.2. *The transverse intersection of $v(Z)$ and $v = Z$ in (Z, v) space.*

by solving $Z_0 = v(Z_0)$, which yields

$$Z_0 = \left(\frac{p^2}{12\sqrt{2}c^4} \right)^{1/3}. \tag{3.4}$$

Therefore (3.4) determines the unique value Z_0 that satisfies (2.7). Transversality follows since the slope of T_0 at the point Z_0 is not equal to one, which is the slope of $v = Z$. □

Remark 3.2. While we use the closed form of $u(\xi)$ to obtain transversality, it will be clear from the estimates for the Gierer–Meinhardt equations in section 6 that the transversality can be obtained in the absence of a closed form solution.

Lemma 3.1 is sufficient to prove that $B_0 \cdot 1$ transversely intersects B_1, the restriction of B_R to the line $v = Z$. Since v changes only along the inner solution, in the (Z, v) plane, the curves T_0 and $B_0 \cdot 1$ are identical. Thus transversality follows from the lemma and is stated in the following corollary.

COROLLARY 3.3. *The curve $B_0 \cdot 1$ transversely intersects B_1 at $v = Z_0$ in (Z, v) space.*

3.1. The argument for $0 < \epsilon \ll 1$. A region of an invariant manifold M_0 is normally hyperbolic if all of the eigenvalues corresponding to eigenvectors normal to the manifold are bounded away from the imaginary axis. Fenichel [8] showed that a normally hyperbolic invariant manifold persists under perturbations. Moreover, the perturbed manifold M_ϵ is $O(\epsilon)$ close to M_0 and also retains this hyperbolic structure.

Finally he showed that the flow on M_ϵ is $O(\epsilon)$. For any Z, recall that $(u, w) = (0, 0)$ is a saddle point for the inner flow and that u and w are restricted to these values on B_0 and B_1. It follows then that the boundary manifolds B_0 and B_1 are normally hyperbolic. Therefore, they perturb to nearby manifolds B_0^ϵ and B_1^ϵ. We need to show that the forward evolution of B_0^ϵ transversely intersects B_1^ϵ at $x = 1$. Instead of making the calculation at $x = 1$, it is more convenient to check this intersection at some intermediate value $x = a$ by flowing B_1^ϵ backwards in space. We will assume that $x = a$ is in some sufficiently small deleted neighborhood of $x = 1/2$.

We need one more critical result concerning the singular flows. Figure 3.1 gives a picture of how the singular flows evolve, but it is deceptive in that it does not fully reveal the transversality that exists in the equations. Indeed, in Figure 3.1, it appears that there is no transversality due to the inner equations. While it is certainly true that the manifold leaving J_0 does not transversely intersect the manifold approaching T_0, we are not actually interested in T_0.

Instead, flow B_1 backwards to $x = 1/2$ under the outer flow. Define a new consistency curve by

$$T_c = \{(u, w, v, Z, x) : u = w = 0, v = Z, x = 1/2\}.$$

Consider once again the inner equations (2.5). Let $W^u(J_0)$ denote the two-dimensional center-unstable manifold of J_0 composed of the union of the one-dimensional unstable manifolds of the critical point $(0, 0)$ over different values of Z. Similarly, let $W^s(T_c)$ denote the two-dimensional center-stable manifold of T_c. Both of these manifolds exist since $(0, 0)$ is a hyperbolic critical point. Note that $W^u(J_0) \to T_0$ as $\xi \to \infty$, and that T_c projected into (Z, v) space is the line $v = Z$. The following is a corollary of Lemma 3.1 and Corollary 3.3.

COROLLARY 3.4. *The manifold $W^u(J_0)$ transversely intersects $W^s(T_c)$ in (u,w,Z) space at $Z = Z_0$.*

The corollary shows that the inner flow induces a transverse intersection of *the* manifolds needed to actually construct the 1-pulse solution. This transversality encodes the consistency condition. Finally, for later use, let $M_0 = \cup_{y\in[0,1/2]} B_0 \cdot y$ and $\mathcal{M}_0 = \cup_{y\in[1/2,1]} T_c \cdot y$, under the outer flow (2.5).

Tin, Kopell, and Jones [22] give conditions for general boundary value problems for which the existence of a singular solution implies the existence of an actual solution for ϵ small. Consistent with the major simplifications offered by geometric singular perturbation theory, these conditions are on the $\epsilon = 0$ singular manifolds. Thus the verification of these conditions occurs in lower-dimensional reduced settings. Tin, Kopell, and Jones' work is based on the exchange lemma of Jones and Kopell [16], which itself relies on Fenichel's invariant manifold theory. The hypotheses (H1)–(H3) below, which provide sufficient conditions to prove the existence of an actual solution for ϵ sufficiently small, are all based on transversality at $\epsilon = 0$ [22]. Stated in notation adapted for this paper, they are the following:

(H1) The outer flow on M_0 transversely intersects J_0.

(H2) $W^u(J_0)$ transversely intersects $W^s(T_c)$ in (u, w, Z) space.

(H3) The outer flow on $\mathcal{M}_0$ transversely intersects T_c.

For our situation, (H1) and (H3) are trivial to verify as can be seen in Figure 3.1. Note that these transversality calculations need to be verified in only a two-dimensional ambient space. Hypothesis (H2) follows directly from Corollary 3.4. Although the ambient space for this intersection is three-dimensional, the needed calculation occurs in a two-dimensional space.

Tin, Kopell, and Jones' results imply the existence of an actual solution for ϵ sufficiently small for the following reason. Let $\mathcal{B}_0^\epsilon$ and $\mathcal{B}_1^\epsilon$ denote the manifolds obtained by flowing B_0^ϵ forward and B_1^ϵ backward under (2.2), respectively. The perturbed manifold $\mathcal{B}_0^\epsilon$ is $O(\epsilon)$ close to M_0, up to a neighborhood of $x = 1/2$. Similarly $\mathcal{B}_1^\epsilon$ is $O(\epsilon)$ close to $\mathcal{M}_0$. Based on the exchange lemma [16], Tin, Kopell, and Jones show that when $\mathcal{B}_0^\epsilon$ and $\mathcal{B}_1^\epsilon$ veer away from these outer manifolds, they are $C^1 - O(\epsilon)$ close to $W^u(J_0)$ and $W^s(T_c)$, respectively. Thus not only are the perturbed manifolds $O(\epsilon)$ close to relevant singular manifolds, so are their tangent spaces. Since transversality is determined by the behavior of tangent spaces, the C^1 closeness is important. Therefore, since $W^u(J_0)$ and $W^s(T_c)$ intersect transversely at Z_0 independent of ϵ, the $C^1 - O(\epsilon)$ closeness of the perturbed manifolds $\mathcal{B}_0^\epsilon$ and $\mathcal{B}_1^\epsilon$ to these manifolds implies that they also intersect transversely for some $Z_\star$ $O(\epsilon)$ close to Z_0.

Therefore, we have shown that $B_0^\epsilon \cdot 1$ transversely intersects B_1^ϵ for ϵ sufficiently small. In $\mathbf{R}^2$, the unique point of intersection of these two curves determines $Z_\star$. This value of $Z_\star$ determines $I_\star$ which is then used in (1.2) to obtain $U_1(x)$. The value of $U_1(0) = U_1(1)$ can then be found. To determine U_{max}, we use the Hamiltonian associated with the inner equations (2.6):

$$H(u,w) = w^2/2 - u^2 + \frac{pu^3}{3c^2Z_0^2}. \tag{3.5}$$

Since the value of H is conserved along trajectories, and, in particular, along the homoclinic of Figure 2.1, we can solve $H(0,0) = H(u_{max}, 0)$. This yields $u_{max} = 3c^2Z_0^2/p$. Finally, rescaling (u, v, w, Z) back to the original (U, W, V, I) variables and using Fenichel [8], the values U_{max} and $I_\star$ stated in Theorem 1.1 are obtained.

Remark 3.5. The geometric argument presented above fails to uniquely pick out the point at which the 1-pulse is centered. Indeed, it is only by using symmetry of the 1-pulse solution that we define the inner equations at $x = 1/2$ and not at some other point in the domain. See Ward [23] for a thorough discussion on this indeterminacy. As he shows, the metastability of solutions is a direct consequence of it.

Remark 3.6. In (1.2), we have not included the $O(\beta)$ terms which actually appear in Kriegsmann's model [18]. The geometric results presented here can be extended to the case where these terms are included in the vector field provided that β is $O(\epsilon^{2+\alpha})$, $\alpha > 0$. The reason why this restriction is needed is precisely because of characteristic (C3) of Kriegsmann's solution. On the interior layer, $U \to \infty$. Thus the included quartic U^4 would be the dominant term of the rescaled equations (2.1) unless β is $O(\epsilon^{2+\alpha})$. As a result, the existence and transverse intersection of $B_0^\epsilon \cdot 1$ with B_1^ϵ persists provided β is small enough.

3.2. Existence of n-pulse solutions. Equation (1.2) also admits n-pulse solutions for small values of ϵ. These solutions contain n equal, local maxima and are also symmetric about $x = 1/2$. A similar geometric argument as above could be given to construct these solutions. Instead, we present a simple rescaling argument which exploits the symmetry requirements of the solutions.[1]

Denote the symmetric 1-pulse solution by $\Phi_1(x)$. This solution satisfies

$$\epsilon^2 \Phi_{1_{xx}} + \frac{pf(\Phi_1)}{1 + c^2(\int_0^1 f(\Phi_1)\,dx)^2} - h(\Phi_1) = 0,$$

[1] I thank Gregory Kriegsmann for suggesting the rescaling argument to me.

$$\Phi_1'(0) = \Phi_1'(1) = 0.$$

Next define

$$\Phi_2(x) = \begin{cases} \Phi_1(x), & 0 < x < 1, \\ \Phi_1(x-1), & 1 < x < 2. \end{cases}$$

Then $\Phi_2(2x)$ satisfies

$$\frac{\epsilon^2}{4}\Phi_{2_{xx}} + \frac{pf(\Phi_2)}{1 + c^2(\int_0^1 f(\Phi_2)\, dx)^2} - h(\Phi_2) = 0,$$

$$\Phi_2'(0) = \Phi_2'(1) = 0.$$

Thus

$$\Phi_2(2x) = \begin{cases} \Phi_1(2x), & 0 < x < 1/2, \\ \Phi_1(2x-1), & 1/2 < x < 1, \end{cases}$$

and $\Phi_2(2x)$ has two local maximum at $x = 1/4$ and $x = 3/4$. Thus $\Phi_2(2x)$ is a symmetric 2-pulse solution of (1.2) which has been obtained by reflecting and rescaling the 1-pulse solution $\Phi_1(x)$. The diffusion constant is half that of the corresponding 1-pulse solution. Now continue the process to obtain a symmetric 4-pulse solution and so on. It is seen that a symmetric $\Phi_{2^m}(2^m x)$ solution exists which satisfies (1.2).

There is nothing special about base 2. In fact, defining

$$\Phi_3(x) = \begin{cases} \Phi_1(x), & 0 < x < 1, \\ \Phi_1(x-1), & 1 < x < 2, \\ \Phi_1(x-2), & 2 < x < 3, \end{cases}$$

we see that $\Phi_3(3x)$ satisfies (1.2) and is a symmetric 3-pulse solution. Continuing in this fashion, it is easy to obtain a symmetric n-pulse solution for any value of n.

In closing, we do not obtain local uniqueness of these solutions by this rescaling method. However, in the next section, we show that any n-pulse solutions for $n \geq 2$ are unstable, so this lack of information concerning local uniqueness is not so important.

4. Stability. In this section, we prove Theorem 1.2. In particular, we show that the 1-pulse solution constructed above is metastable, with its principal eigenvalue being exponentially small in ϵ. The n-pulse solutions, however, are unstable, with principal eigenvalues bounded away from the origin as $\epsilon \to 0$.

Our analysis relies on two key ingredients. The first is an oscillation theorem for the nonlocal eigenfunctions and their corresponding eigenvalues. This theorem is similar to the one found in Bose and Kriegsmann [1], but is established here under more general conditions. We need a different argument to prove it due to the choice of $f(u)$. The second ingredient will be an analysis of Ward [23], which shows the existence of an exponentially small eigenvalue for equations of the type that we consider.

The oscillation theorem that we prove below is quite general. It holds for a large class of nonlinearities, provided that the underlying pulse solution is symmetric about the midpoint of the domain of interest. In particular, the theorem also holds outside of the singular perturbation parameter regime. However, using the oscillation theorem to prove the metastability of the 1-pulse uses the singular structure in two important ways. First, it is needed to establish the existence of an exponentially small eigenvalue. Second, it is used to rule out the existence of an eigenfunction of strictly one sign. For local equations, it is well known that pulse solutions have O(1) with respect to ϵ unstable principal eigenvalues. The corresponding eigenfunction, after normalization, is strictly positive. Using our oscillation theorem and information about the structure of solutions as $\epsilon \to 0$, we show that no such eigenfunction can exist for the nonlocal problem. Thus, the nonlocal term can be viewed as stabilizing an unstable solution of the local problem.

In the standard manner, assume $U(x,t) = U_1(x) + \phi(x)e^{-\lambda t}$ and linearize (1.1) about U_1 to obtain the following nonlocal eigenvalue problem:

$$\epsilon^2\phi'' + (A(x)+\lambda)\phi = B(x)\int_0^1 C(x)\phi\,dx, \tag{4.1}$$

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

where

$$A(x) = -2 + \frac{2pU_1}{1+c^2(1+I_\star)^2},\ B(x) = \frac{2pc^2(1+I_\star)(1+U_1^2)}{(1+c^2(1+I_\star)^2)^2},\ C(x) = \ 2U_1. \tag{4.3}$$

Let L_1 be the linear operator associated with (4.1), which is given by

$$L_1\phi = -\epsilon^2\phi'' - A(x)\phi + B(x)\int_0^1 C(x)\phi\ dx. \tag{4.4}$$

Denote the spectrum of L_1 by $\sigma(L_1)$. If $Re\ \sigma(L_1) > 0$, then U_1 will be an asymptotically stable solution of (1.1) [3]. In [1], $f(u) = e^{c_1 u}$ which implies that L$_1$ is self-adjoint. Now, since $f(u) = 1+u^2$, L_1 is not a self-adjoint operator. Thus there is no a priori guarantee that the eigenvalues of L_1 are real. We show below that due to the symmetry of U_1, the eigenvalues of L_1 must in fact be real.

The spectrum of L_1 can be related to the eigenvalues of the following Sturm–Liouville equation:

$$\epsilon^2\psi'' + (A(x)+\nu)\psi = 0, \tag{4.5}$$

$$\psi'(0) = \psi'(1) = 0. \tag{4.6}$$

Denote by L_0 the corresponding linear operator. For L_0, there exists a sequence of eigenvalues $\{\nu_n\}$ such that $\nu_0 < \nu_1 < \nu_2 \dots$ and corresponding eigenfunctions $\{\psi_n\}$ such that each eigenfunction has exactly n interior zeros [4]. Due to the fact that U_1 is symmetric about $x = 1/2$, it turns out that the eigenfunctions $\{\psi_n\}$ break up into two subsets: $\{\psi_{2n}\}$ which are even about $x = 1/2$ and $\{\psi_{2n+1}\}$ which are odd about $x = 1/2$. This occurs because $A(x)$, $B(x)$, and $C(x)$ all must be even about $x = 1/2$. As a result, note that

$$\int_0^1 C(x)\psi_{2n+1}\ dx = 0. \tag{4.7}$$

Therefore the odd local eigenpairs $(\nu_{2n+1}, \psi_{2n+1})$ also turn out to be nonlocal eigenpairs. This observation forms the basis for the following oscillation theorem for the nonlocal eigenvalues and eigenfunctions.

Oscillation Theorem. *Let λ be a nonlocal eigenvalue of L_1 with corresponding eigenfunction ϕ. For $n \geq 1$,*

(a) *$\lambda = \nu_{2n-1}$ if and only if $\phi = \psi_{2n-1}$ has $2n - 1$ interior zeros.*

(b) *$\nu_{2n-1} < \lambda < \nu_{2n+1}$ if and only if ϕ has $2n$ interior zeros.*

(c) *Every interval (ν_{2n-1}, ν_{2n+1}) contains exactly one nonlocal eigenvalue except possibly one such interval which may contain at most two nonlocal eigenvalues.*

Remark 4.1. The oscillation theorem, along with Lemma 4.2 below, gives a complete description of $\sigma(L_1)$. Moreover, as will be apparent from the proof, the exact forms of the nonlinearities $f(u)$ and $h(u)$ are never used. As a result, the oscillation theorem holds quite generally.

Proof. Part (a) of the theorem is obvious. Part (b) is proved by Bose and Kriegsmann in [1]. To prove part (c), we need a different argument than in [1]. There, because $f(u) = e^{c_1 u}$, we were able to explicitly show that in the interval (ν_{2n-2}, ν_{2n}), there exists exactly one nonlocal eigenvalue. In the present situation, we have no information about the interval (ν_{2n-2}, ν_{2n}). We show, however, that the symmetry of U_1 forces part (c) to hold.

We need the following result.

Lemma 4.2. *The eigenvalues of L_1 are strictly real.*

Proof. Following Freitas [10], we introduce the parameter δ to define

$$L_\delta \phi = L_0 \phi + \delta B(x) \int_0^1 C(x)\phi \, dx. \tag{4.8}$$

Note that $\delta = 1$ yields L_1 as defined in (4.4) and $\delta = 0$ yields L_0. Freitas shows that the eigenvalues $\lambda(\delta)$ of L_δ vary continuously with δ. If $\lambda \in \sigma(L_\delta)$ for all δ, then Freitas calls the eigenvalue a fixed eigenvalue, otherwise the eigenvalue is a moving eigenvalue. Generically, as δ is varied from zero, some of the eigenvalues $\lambda(\delta)$ begin to move along the real axis. Freitas shows that eigenvalues cannot suddenly appear or disappear as δ is varied. Moreover, he shows that the only way complex conjugate eigenvalues can be created is if for some value of δ, two moving eigenvalues collide while traveling in opposite directions. Then for some nonzero δ interval, perhaps semi-infinite, they remain complex. Finally, Freitas shows that moving eigenvalues which are traveling in the same direction cannot cross over one another. We see here that the odd subscripted eigenvalues $\lambda_{2n+1}(\delta) = \nu_{2n+1}$ are fixed eigenvalues and that the even ones $\lambda_{2n}(\delta)$, where $\lambda_{2n}(0) = \nu_{2n}$, are moving eigenvalues.

In [1], the Prüfer transformation, $\tan \epsilon\phi'/\phi$, is used to show that for $\delta = 1$, there must exist a (moving) nonlocal eigenvalue in every subinterval (ν_{2n-1}, ν_{2n+1}). That same argument actually shows that this must be true for all δ. This observation is sufficient to rule out the existence of complex conjugate eigenvalues. We argue by contradiction. Suppose that there exists $\hat{\delta}$ at which $\lambda_0(\hat{\delta})$ and $\lambda_2(\hat{\delta})$ collide and become complex. Then since there must exist a real nonlocal eigenvalue in (ν_1, ν_3), it follows that $\lambda_4(\hat{\delta}) \in (\nu_1, \nu_3)$. In particular, there exists $\delta_4 < \hat{\delta}$ such that $\lambda_4(\delta_4) = \nu_3$. Since for any value of δ, (ν_3, ν_5) must also contain a real nonlocal eigenvalue, it follows that $\lambda_6(\delta_4) \in (\nu_3, \nu_5)$. Thus there exists a $\delta_6 < \delta_4 < \hat{\delta}$ such that $\lambda_6(\delta_6) = \nu_5$. Continuing by induction, there exists a monotone decreasing sequence $\{\delta_{2n}\}$ such that $\lambda(\delta_{2n}) = \nu_{2n-1}$ for $n \geq 2$.

The sequence $\{\delta_{2n}\}$ denotes the "crossing times" at which a moving eigenvalue crosses over the fixed eigenvalue immediately to its left. Since the sequence $\{\delta_{2n}\}$ is monotone decreasing, let us assume that $\lim_{n\to\infty} \delta_{2n} = \delta^\star$, where $\delta^\star \geq 0$. If $\delta^\star < 0$, then we are done. At $\delta = \delta^\star$, no moving eigenvalues can have crossed fixed ones.

But for $\delta = \delta^\star + \alpha$, where $\alpha > 0$ but arbitrarily small, an *infinite* number of moving eigenvalues must have crossed over fixed ones. The following lemma, which is a direct consequence of Proposition 3.8 in Freitas [10], shows that this situation cannot occur.

LEMMA 4.3. *For any fixed $\delta > 0$, only a finite number of moving eigenvalues can have crossed fixed ones.*

Proof. We sketch the proof here. The details of the proof can be found in Freitas [10]. Let $\eta \in \rho(L_0)$, where ρ is the resolvent set of L_0. Let $R(\eta, L_0)$ denote the resolvent of L_0. Using a general result of Kato [17], Freitas shows that if $\delta \int_0^1 B(x)\phi_{2n}\, dx \int_0^1 C(x)\phi_{2n}\, dx \cdot ||R(\eta, L_0)|| < 1$, then $\eta \in \rho(L_\delta)$. Moreover, again using Kato, he shows that $||R(\eta, L_0)|| = (\text{dist}(\eta, \sigma(L_0)))^{-1}$. Therefore if

$$\delta \int_0^1 B(x)\phi_{2n}\, dx \int_0^1 C(x)\phi_{2n}\, dx < \text{dist}(\eta, \sigma(L_0)), \tag{4.9}$$

then $\eta \in \rho(L_\delta)$, i.e., $\eta \notin \sigma(L_\delta)$. Freitas shows that since $\nu_{2n+1} - \nu_{2n} \to \infty$ as $n \to \infty$, the $\text{dist}(\eta, \sigma(L_0))$ can be made arbitrarily large for sufficiently large n. Since $B(x)$ and $C(x)$ are functions associated with the linearization about U_1 and are independent of n and since ϕ_{2n} is increasingly oscillatory as $n \to \infty$, $\int_0^1 B(x)\phi_{2n}\, dx \int_0^1 C(x)\phi_{2n}\, dx$ is bounded as $n \to \infty$. Thus, for any fixed value of δ, if n is sufficiently large, (4.9) holds. This means that if δ is fixed, if n is sufficiently large, and if $\text{Re}\ \eta \in (\nu_{2n-1}, \nu_{2n})$, then $\lambda_{2n}(\delta) > \text{Re}\ \eta$. Therefore only a finite number of moving eigenvalues can have crossed fixed ones. □

Lemma 4.3 stands in direct contradiction to the construction of our sequence $\{\delta_{2n}\}$. Recall, that the existence of such a sequence is a necessary condition for complex conjugate eigenvalues to exist. Thus, we conclude that no complex eigenvalues can exist for $\delta > 0$ and, in particular, for $\delta = 1$. □

Using Lemma 4.2, we can now prove part (c) of the oscillation theorem. From the proof of Lemma 4.2, a moving eigenvalue $\lambda_{2n}(\delta)$ can only cross over the fixed eigenvalue ν_{2n+1} that lies immediately to its right. Moreover, for all δ every interval (ν_{2n-1}, ν_{2n+1}) must contain at least one nonlocal eigenvalue. Thus if $\lambda_0(\delta) < \nu_1$ for all δ, then for $n \geq 1$, every interval (ν_{2n-1}, ν_{2n+1}) must contain exactly one nonlocal eigenvalue. Next suppose $\lambda_0(1)$ and $\lambda_2(1)$ are both in (ν_1, ν_3), then for $n \geq 2$, (ν_{2n-1}, ν_{2n+1}) contains exactly one nonlocal eigenvalue. If (ν_3, ν_5) contains two nonlocal eigenvalues, then for $n \geq 3$, (ν_{2n-1}, ν_{2n+1}) contains exactly one nonlocal eigenvalue. It also implies that $\lambda_2(1)$ and $\lambda_4(1)$ are elements of (ν_3, ν_5). The eigenvalue $\lambda_0(1)$ must remain in (ν_1, ν_3). Continuing in this manner, we see that at most one interval (ν_{2n-1}, ν_{2n+1}) contains two nonlocal eigenvalues. □

See [1] for further remarks concerning the oscillation theorem.

4.1. Analysis for the 1-pulse U_1.

The specific analysis to prove the metastability of the 1-pulse is very similar to that in Bose and Kriegsmann [1]. We sketch only the results.

LEMMA 4.4. *The nonlocal eigenvalue $\lambda_1 = \nu_1 < 0$ but is exponentially small in ϵ. Moreover, it is the principal eigenvalue of L_1.*

Proof. Consider (4.1) with Dirichlet boundary conditions. The derivative of the 1-pulse, $\zeta = U_1'$, is an eigenfunction of this equation with a corresponding eigenvalue of zero. Thus ζ satisfies

$$\epsilon^2 \zeta'' + A(x)\zeta = 0, \tag{4.10}$$

$$\zeta(0) = \zeta(1) = 0. \tag{4.11}$$

Next, consider (4.1) for the eigenpair (ϕ_1, λ_1). A standard trick is to multiply (4.1) by ζ, (4.10) by ϕ_1, subtract the two ensuing equations, and then integrate by parts. Doing so yields that λ_1 is $O(\epsilon^2 U_1''(1))$ and negative; see [1] for specifics. Ward [23] shows that $\epsilon^2 U_1''(1)$ is $O(e^{-a/\epsilon})$, with $a > 0$, for the class of pulse like solutions under consideration here. See [23] for a detailed derivation of this result. Thus λ_1 is exponentially small in ϵ.

We next show that L_1 has no eigenfunction of strictly one sign. Thus, part (b) of the oscillation theorem will imply that λ_1 is the principal eigenvalue. We argue by contradiction. Assume that there exists $\phi_0 > 0$ for all $x \in [0,1]$ which satisfies (4.1–4.2). Let λ_0 be its associated eigenvalue. Let $J = \int_0^1 2U_1\phi_0 \, dx$. Note that $J > 0$. Integrating (4.1) on $[0,1]$ yields

$$(\lambda_0 - 2)\int_0^1 \phi_0 \, dx + \frac{pJ}{1 + c^2(1+I_\star)^2} = \frac{2pc^2 J(1+I_\star)^2}{(1+c^2(1+I_\star)^2)^2}. \tag{4.12}$$

Rearranging and factoring common terms yields

$$(\lambda_0 - 2)\int_0^1 \phi_0 \, dx = \frac{pJ}{(1+c^2(1+I_\star)^2)^2}\left[c^2(1+I_\star)^2 - 1\right]. \tag{4.13}$$

From Theorem 1.1, we know that $I_\star \to \infty$ as $\epsilon \to 0$. Thus for ϵ sufficiently small, the right-hand side of (4.13) is strictly positive. Since $\int_0^1 \phi_0 \, dx > 0$ by assumption, this implies $\lambda_0 > 2$. This however contradicts part (b) of the oscillation theorem as λ_0 cannot by greater than ν_1. Thus there is no eigenfunction of strictly one sign. □

4.2. Analysis for the n-pulse solutions U_n. The eigenvalue equation for the n-pulse is identical to (4.1) except that U_1 is replaced by U_n. Recall that the n-pulse solutions were generated using the symmetry of the 1-pulse. As a result, information about the principal eigenvalue for the linearization around these pulses can be obtained from the 1-pulse. Assume first that $n = 2$. The analysis for $n \geq 3$ is identical to this case. As shown in [1], it turns out that λ_1 is $O(\epsilon^2 U_1''(1/2))$. Notice the difference between this relationship and that for the 1-pulse presented above. Since $U_1(1/2) = U_{max}$, by (1.2),

$$\epsilon^2 U_1''(1/2) = 2U_{max} - \frac{p(1+U_{max}^2)}{1+c^2(1+I_\star)^2}. \tag{4.14}$$

Using the estimates from Theorem 1.1, we obtain

$$\epsilon^2 U_1''(1/2) = \frac{3c^2 I_\star^2}{p}\left(2 - \frac{3c^2 I_\star^2}{1+c^2(1+I_\star)^2}\right). \tag{4.15}$$

As ϵ tends to zero, the term in parentheses tends to negative one, while the factor multiplying this tends to infinity. Thus the right-hand side tends to negative infinity. Therefore $\lambda_1 \to -\infty$ as $\epsilon \to 0$. This shows that the principal eigenvalue of the 2-pulse (and analogously the n-pulses) is unstable and bounded away from the origin for ϵ sufficiently small.

5. Numerical simulations. In this section we present a few numerical simulations. We solved the time dependent equation (1.1) using an implicit Crank–Nicholson scheme as in [1, 18]. For all simulations we ran the code for around 20,000 time steps, which corresponds to about 12 seconds. To create a 1-pulse solution, we evolved an

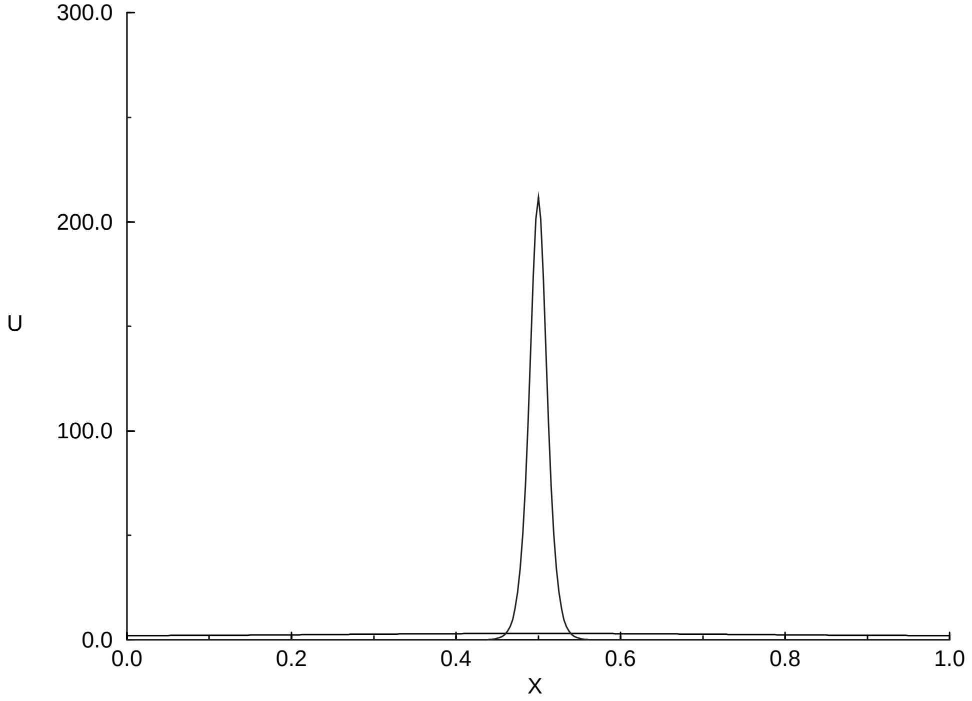


FIG. 5.1. *The* 1*-pulse solution with* $\epsilon = 0.01$*,* $p = 1.0$*,* $c = 0.01$*.*

initial condition that had one small local maxima at $x = 1/2$; see Figure 5.1. To test whether our analytic predictions of U_{max} and $I_\star$ are reasonable, we ran the code with several different choices of p and c and a few different values of ϵ and numerically calculated these values. In Table 5.1, we show results for $p = 1.0$. Different values for ϵ and c are presented. Kriegsmann [18] uses $\epsilon = 0.01$ and $c = 0.01$ in his simulations, which are both physically relevant parameter values. As can be seen, the numerical values agree closely with the theoretical predictions, thus implying a consistency between our numerical and analytic results.

TABLE 5.1
A comparison of analytically predicted and numerically calculated values of U_{max} *and* $I_\star$*.*

ϵ	c	Numerical value of U_{max}	Analytic value of U_{max}	Numerical value of $I_\star$	Analytic value of $I_\star$
0.01	0.01	211.77	210.85	831.41	838.37
0.03	0.01	100.32	101.37	568.52	581.29
0.01	0.1	43.96	45.42	36.89	38.9

Given such close agreement, we further pursued the ramifications of the metastability of the 1-pulse by evolving different initial conditions. Figure 5.2 shows the evolution of a symmetric 2-bump perturbation of a homogeneous solution for $\epsilon = 0.01$. We show only the initial condition and the ensuing 1-pulse. Notice that it is not centered at the midpoint of the domain, so it is not a steady-state solution. When we let the code run for much longer times, the 1-pulse remained fixed. This is a manifestation of the metastability of the solution. This non–steady-state 1-pulse is moving exponentially slowly towards one of the boundaries. We do not know why the pulse

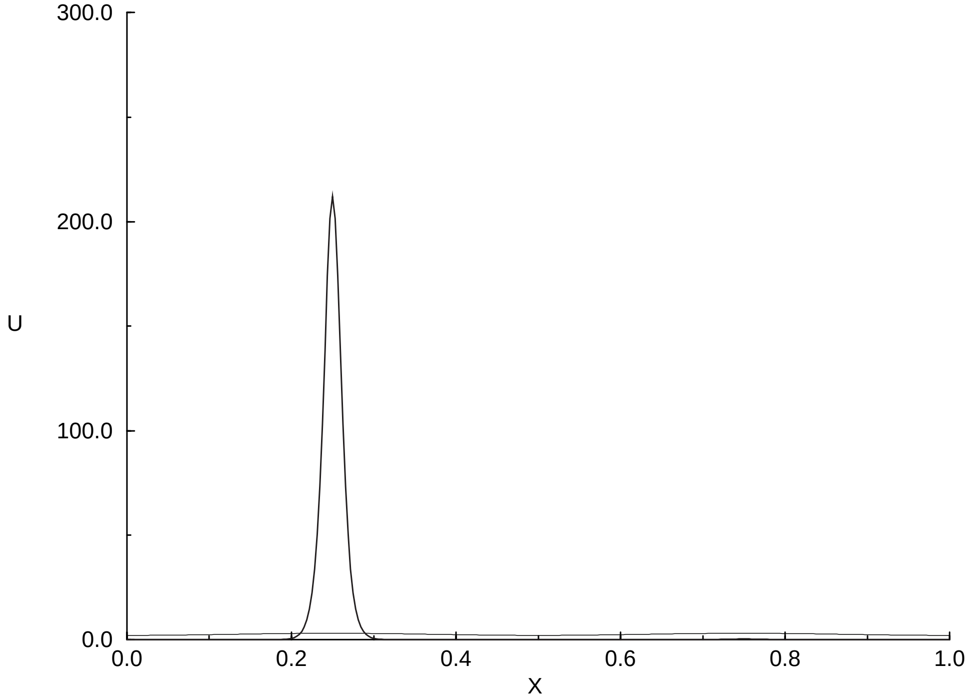


FIG. 5.2. *An exponentially slowly moving* 1*-pulse with* $\epsilon = 0.01$*,* $p = 1.0$*,* $c = 0.01$*.*

forms at $x = 1/4$ and not at $x = 3/4$. Using other initial conditions, we can generate a pulse at any location in the domain (simulations not shown). The exponentially small eigenvalue introduces a near translational invariance, so this is not surprising. Incidentally, it is possible to obtain a 1-pulse centered at $x = 1/2$ by providing an initial condition with a sufficiently large local maximum at $x = 1/2$. These results are consistent with [1].

Finally, using the rescaling argument in section 3.2, we can obtain so-called 1/2-pulse steady-state solutions. These solutions are close to zero valued on most of $(0, 1)$, but contain a layer near one of the boundaries where the value of U grows inversely to some power of ϵ. We call them 1/2-pulses since they look like the 1-pulse solution over the interval $[0, 1/2]$ (or alternatively, $[1/2, 1]$). We also observed such solutions numerically, which leads us to believe that they are stable solutions. We do not, however, pursue their stability here.

6. Application to the Gierer–Meinhardt equations.

The Gierer–Meinhardt equations [13] arise in biological pattern formation. They fall in the general class of activator-inhibitor systems. They can be written in nondimensional form in one spatial dimension as

$$U_t = \epsilon^2 U_{xx} - U + \frac{U^p}{H^q}, \tag{6.1}$$

$$\tau H_t = D_H H_{xx} - \mu H + \frac{U^m}{H^s}, \tag{6.2}$$

$$U_x(0,t) = U_x(1,t) = 0,\ H_x(0,t) = H_x(1,t) = 0. \tag{6.3}$$

The variable U represents the activator concentration and H represents the inhibitor concentration. Here $p > 1$, $q, m > 0$, and $s \geq 0$. These equations were studied numerically in [13] and localized pulses were observed. Recently, Iron and Ward [14] used numerical and asymptotic techniques to demonstrate the metastability of the observed 1-pulse solution of a nonlocal reduction of (6.1) for $x \in \Omega$, where Ω is a closed subset of $\mathbf{R^N}$. They derived a nonlocal reaction-diffusion equation which was valid in the limit as $\tau \to 0$ and $D_H \to \infty$. They did not rigorously construct solutions nor prove their stability. Here, for $x \in [0, 1]$ we derive a slightly different scaled version of the nonlocal equation that appears in [14] and then construct a 1-pulse solution. As before n-pulses can be obtained by the rescaling argument. Moreover, the metastability analysis will be exactly as in section 4.

As $D_H \to \infty$, H becomes spatially homogeneous. We then integrate (6.2) from zero to one and set $\tau = 0$ to obtain the following algebraic equation:

$$\mu H^{s+1} = \int_0^1 U^m \, dx. \tag{6.4}$$

Setting $\mu = 1$, for convenience, we obtain the following scalar nonlocal reaction-diffusion equation:

$$U_t = \epsilon^2 U_{xx} - U + \frac{U^p}{(\int_0^1 U^m \, dx)^{q/(s+1)}}, \tag{6.5}$$

$$U_x(0, t) = U_x(1, t) = 0.$$

Iron and Ward first scale the system (6.1–6.2) to reflect that the height of the pulse goes to infinity as ϵ tends to zero. They then integrate (6.2). The difference is that they have a factor of $(1/\epsilon)^{q/(s+1)}$ multiplying the nonlocal term.

As before, we rewrite (6.5) as a system of first-order equations using $I = \int_0^1 U^m \, dx$ and the auxiliary variable $V(x) = \int_0^x U^m \, dx$. Set $n = q/(s + 1)$.

$$\begin{aligned} \epsilon U' &= W, & V' &= U^m, \\ \epsilon W' &= U - \tfrac{U^p}{I^n}, & I' &= 0. \end{aligned} \tag{6.6}$$

Introducing the scalings $u = \epsilon^a U$, $w = \epsilon^a W$, $v = \epsilon^b V$, and $Z = \epsilon^b I$ as before and balancing terms yields the following values for a and b:

$$a = \frac{q}{qm - p(s + 1)}, \qquad b = \frac{p(s + 1)}{qm - p(s + 1)}.$$

Thus we require $qm > p(s + 1)$, which is consistent with [13, 14]. This yields

$$\begin{aligned} \epsilon u' &= w, & \epsilon v' &= u^m, \\ \epsilon w' &= u - \tfrac{u^p}{Z^n}, & Z' &= 0. \end{aligned} \tag{6.7}$$

If we rescale in a neighborhood of $x = 1/2$ using $\xi = (x - 1/2)/\epsilon$, then the ensuing system has one saddle point and one center point as before. Similarly, there exists a homoclinic solution connecting the saddle to itself. Using the change of variable $y = x$ and appending $dx/dy = 1$, the boundary manifolds, jump off, and touch down curves are defined exactly as before. Consistency is again checked on the inner equations.

As before, the outer flow transversely intersects J_0 and T_c, so (H1) and (H3) are satisfied. To verify (H2), we must check transversality of $B_0 \cdot 1$ with B_1. We use the inner equations associated with (6.7) which are

$$\begin{aligned} \dot{u} &= w, & \dot{v} &= u^m, \\ \dot{w} &= u - \tfrac{u^p}{Z^n}, & \dot{Z} &= 0. \end{aligned} \tag{6.8}$$

The first two equations have Hamiltonian given by

$$H(u,w) = w^2/2 - u^2/2 + \frac{u^{p+1}}{(p+1)Z^n}. \tag{6.9}$$

The maximum value of u is given by $u_{max} = ((p+1)Z^n/2)^{1/(p-1)}$. Recall that the curve $v(Z) = \int_{-\infty}^{\infty} u^m(\xi, Z)\, d\xi$ is the projection of the touch down curve T_0 on the (Z, v) plane. We must show that this curve transversely intersects the line $v = Z$. To do this we obtain estimates for $\int_{-\infty}^{\infty} u^m\, d\xi$ for $Z \ll 1$ and $Z \gg 1$ and use the intermediate value theorem.

Recall that the homoclinic solution of the first two equations of (6.8) corresponds to the pulse. The value of the Hamiltonian on this solution is 0. Using (6.9), the integral of interest can be rewritten as

$$\int_{-\infty}^{\infty} u^m\, d\xi = 2\int_0^{u_{max}} \frac{u^{m-1}}{(1 - \frac{2u^{p-1}}{(p+1)Z^n})^{1/2}}\, du. \tag{6.10}$$

Moreover,

$$\begin{aligned} \int_0^{u_{max}} \frac{u^{m-1}}{\left(1 - \frac{2u^{p-1}}{(p+1)Z^n}\right)^{1/2}}\, du &> \int_0^{u_{max}} u^{m-1}\, du \\ &= \frac{1}{m}\left(\tfrac{(p+1)Z^n}{2}\right)^{m/(p-1)}. \end{aligned} \tag{6.11}$$

It is easily seen that if $Z \gg 1$, then the right-hand side of the above inequality is greater than Z, since $nm/(p-1) = qm/(p-1)(s+1) > qm/p(s+1) > 1$, by assumption.

Depending on the choices of the parameters m, p, q, and s there are many ways to obtain an upper bound for $\int_{-\infty}^{\infty} u^m\, d\xi$. We assume that $n > 1$ and $p \le m+1$ in the following. Note that $Z < 1$ implies that u_{max} and thus u are less than one. Then

$$\begin{aligned} \int_0^{u_{max}} \frac{u^{m-1}}{\left(1 - \frac{2u^{p-1}}{(p+1)Z^n}\right)^{1/2}}\, du &< \int_0^{u_{max}} \frac{u^{p-2}}{\left(1 - \frac{2u^{p-1}}{(p+1)Z^n}\right)^{1/2}}\, du \\ &= \tfrac{(p+1)Z^n}{(p-1)}. \end{aligned} \tag{6.12}$$

Clearly, if $Z \ll 1$, then the right-hand side of (6.12) is less than Z. For different choices of the parameters m, p, q, and s, other estimates such as (6.12) can be obtained, but we do not pursue them here.

Combining the appropriate estimates above, it is seen that if $Z \ll 1$, then $\int_{-\infty}^{\infty} u^m\, d\xi < Z$, and if $Z \gg 1$, then $\int_{-\infty}^{\infty} u^m\, d\xi > Z$. Thus by the intermediate value theorem, there exists at least one value of Z_0 for which $\int_{-\infty}^{\infty} u^m(\xi, Z_0)\, d\xi = Z_0$. This value is unique as can be inferred from the above estimates. Specifically, for the particular choices $m = p - 1$ and $m = 2p - 2$, it is easy to check by direct integration that $v(Z)$ is $O(Z^n)$ and $O(Z^{2n})$, respectively. For any choice of $m > 2p - 2$, the function $v(Z)$ can be obtained by using a suitable number of integration by parts and eventually reducing to the calculation of an integral of the form (6.10) where $p - 1 \leq m \leq 2p - 2$. In each integration by parts, the boundary terms disappear and an additional factor of Z^n is introduced. By rewriting the relevant integral as the sum of two integrals for the cases $u < 1$ and $u > 1$, if necessary, and using the results obtained from the special cases $m = p-1$ or $m = 2p-2$, we can obtain a lower bound for $v(Z)$ which is at least $O(Z^n)$. This result holds for all $m \geq p - 1$, which is the parameter regime of interest. Since $n > 1$, this clearly shows that $V(Z)$ is not linear in Z. This establishes transversality of T_0 with the line $v = Z$.

The stability analysis for this 1-pulse solution is analogous to what we presented in section 4. The nonexistence of a positive eigenfunction can be established under the further restriction that $p \leq m$. We leave the details to the interested reader. Also, time-dependent simulations of (6.5) using different choices of the parameters, produced metastable 1-pulses as in section 5.

Finally, Iron and Ward [14] discuss the metastability of the 1-pulse in higher spatial dimensions. Moreover, they derive an equation of motion for the metastable pulse and discuss how it interacts with the boundary in which it is enclosed. We refer the interested reader to their work.

7. Discussion. In this paper, we have developed a systematic geometric method to construct spatially localized pulse like solutions for singularly perturbed nonlocal boundary value problems. While we have not stated a general theorem concerning the construction of solutions, it is clear that the procedure outlined above is applicable to a large variety of scalar nonlocal equations. Moreover, the analysis presented above is not restricted to singularly perturbed equations or to the construction of pulse-like solutions. These methods can also be used to construct front-type (heteroclinic) solutions.

We showed how to recast the scalar nonlocal problem as a higher-dimensional local problem. The geometric framework provided above can also extend to higher-dimensional nonlocal systems. In higher dimensions, the jump off and touch down curves may become surfaces, but the abstract description using manifolds accounts easily for this possibility. One aspect of the present low-dimensional analysis that will remain in the higher-dimensional setting will be the transversality of two one-dimensional curves in a two-dimensional ambient space needed to establish the consistency condition (1.3). No matter what the dimensionality of the full system is, this two-dimensional problem will always persist. Thus one of the challenges of any nonlocal analysis is to recast the system into a form where this consistency condition becomes apparent.

A general oscillation theorem was also presented. A sufficient condition for which this theorem holds is the symmetry of the underlying pulse solution. We note, however, that this condition is a very natural one to impose since any steady-state pulse solution of (1.1) must necessarily by symmetric. Modifications of the oscillation theorem should also hold in circumstances where the operator L_1 has an infinite number of fixed eigenvalues. It is hard to give a general example where this may occur, but such a situation may arise when the underlying solution is not strictly of one sign.

We have also proved the metastability of the 1-pulse solution and the instability of n-pulse solutions. As discussed in [1], this type of metastability is qualitatively different than the metastability found in, say, Carr and Pego [2]. There, the authors construct metastable solutions that contain an arbitrarily high number of interior layers. These layers consist of heteroclinic, and not homoclinic, solutions to the appropriate set of inner equations. They show that each interior layer contributes an exponentially small eigenvalue. In our work, we have shown that additional localized pulses contribute $O(1)$ unstable eigenvalues. The nonlocal term can be viewed as being strong enough to remove at most one local unstable eigenvalue.

Finally, the applications considered in this paper are of interest in their own right. Because of the time scales of physical interest, metastability is tantamount to stability. Thus metastability of the 1-pulse of the microwave heating model due to Kriegsmann [18] suggests that localized heating of ceramic materials can be achieved in a stable and reliable manner. The nonlinearities chosen for the present study yield maximum heating rates that are too high and beyond the melting point of the fiber. However, for the nonlinearities used in [1, 18], physically acceptable maximal heating rates are obtained. Other nonlocal models arising in different microwave heating applications are the focus of further research.

Acknowledgment. I offer my most sincere thanks to my colleague Gregory Kriegsmann for introducing me to microwave heating problems and nonlocal equations. This paper is a result of countless discussions that I have had with him on the topic. I also thank him for his unwavering support and encouragement.

REFERENCES

[1] A. Bose and G. A. Kriegsmann, *Stability of localized structures in non-local reaction-diffusion equations*, Methods Appl. Anal., 5, (1998), pp. 351–366.

[2] J. Carr and R. L. Pego, *Metastable patterns in solutions of* $u_t = \epsilon^2 u_{xx} - f(u)$, Comm. Pure Appl. Math., 42 (1989), pp. 523–576.

[3] N. Chafee, *The electric ballast resistor: Homogeneous and nonhomogeneous equilibria*, in Nonlinear Differential Equations: Invariance, Stability and Bifurcations, P. de Mottoni and L. Salvadori, eds., Academic Press, New York, 1981, pp. 97–127.

[4] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[5] A. Doelman, R. A. Gardner, and T. J. Kaper, *Stability analysis of singular patterns in the 1D Gray-Scott model* I: *A matched asymptotic approach*, Phys. D, 122 (1998), pp. 1–36.

[6] A. Doelman, T. J. Kaper, and P. A. Zegeling, *Pattern formation in the one-D Gray-Scott model*, Nonlinearity, 10 (1997), pp. 523–563.

[7] A. Doelman and V. Rottschäfer, *Singularly perturbed and nonlocal modulation equations for systems with interacting instability mechanisms*, J. Nonlinear Sci., 7 (1997), pp. 371–409.

[8] N. Fenichel, *Persistence and smoothness of invariant manifolds of flows*, Indiana Univ. Math. J., 21 (1971/1972), pp. 193–226.

[9] B. Fiedler and P. Poláčik, *Complicated dynamics of scalar reaction diffusion equations with a nonlocal term*, Proc. Roy. Soc. Edinburgh Sect. A, 115 (1990), pp. 167–192.

[10] P. Freitas, *A nonlocal Sturm-Liouville eigenvalue problem*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 169–188.

[11] P. Freitas, *Bifurcation and stability of stationary solutions on nonlocal scalar reaction-diffusion equations*, J. Dynam. Differential Equations, 6 (1994), pp. 613–629.

[12] P. Freitas, *Stability of stationary solutions for a scalar non-local reaction-diffusion equation*, Quart. J. Mech. Appl. Math., 48 (1995), pp. 557–582.

[13] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetika, 12 (1972), pp. 30–39.

[14] D. Iron and M. J. Ward, *A metastable spike solution for a non-local reaction-diffusion model*, SIAM J. Appl. Math, to appear.

[15] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems, Montecatini Terme, 1994, R. Johnson, ed., Springer, Berlin, 1995, pp. 44–118.

[16] C. K. R. T. Jones and N. Kopell, *Tracking invariant manifolds with differential forms in singularly perturbed systems*, J. Differential Equations, 108 (1994), pp. 64–88.

[17] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1980.

[18] G. A. Kriegsmann, *Hot spot formation in microwave heated ceramic fibers*, IMA J. Appl. Math., 59 (1997), pp. 123–148.

[19] A. A. Lacey, *Thermal runaway in a non-local problem modelling Ohmic heating: Part* I: *Model derivation and some special cases*, European J. Appl. Math., 6 (1995), pp. 127–144.

[20] U. Middya and D. Luss, *Impact of global interaction and symmetry on pattern selection and bifurcation*, J. Chem. Phy., 101 (1994), pp. 4688–4696.

[21] Y. L. Tian, *Practices of ultra-rapid sintering of ceramics using single mode applicators*, Cer. Trans., 21 (1991), pp. 283–290.

[22] S. K. Tin, N. Kopell, and C. K. R. T. Jones, *Invariant manifolds and singularly perturbed boundary value problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1558–1576.

[23] M. J. Ward, *Eliminating indeterminacy in singularly perturbed boundary value problems with translation invariant potentials,* Stud. Appl. Math., 87 (1992), pp. 95–135.

SIAM J. MATH. ANAL.
Vol. 31, No. 2, pp. 455–465

NONEXISTENCE OF STAR-SUPPORTED SPLINE BASES[*]

PETER ALFELD[†] AND LARRY L. SCHUMAKER[‡]

Abstract. We consider polynomial spline spaces $\mathcal{S}_d^r(\triangle)$ of degree d and smoothness r defined on triangulations. It is known that for $d \geq 3r+2$, $\mathcal{S}_d^r(\triangle)$ possesses a basis of *star-supported* splines, i.e., splines whose supports are at most the set of triangles surrounding a vertex. Here we extend the theory by showing that for all $d \leq 3r+1$, there exist triangulations for which no such bases exist.

Key words. multivariate splines, piecewise polynomial functions, triangulations

AMS subject classifications. 41A63, 41A15, 65D07

PII. S0036141098342349

1. Introduction.

Given a regular triangulation $\triangle$, let

$$\mathcal{S}_d^r(\triangle) := \{s \in C^r(\Omega) : \; s|_T \in \mathcal{P}_d \text{ for all triangles } T \in \triangle\},$$

where $\mathcal{P}_d$ is the space of polynomials of degree d and Ω is the union of the triangles in $\triangle$. Such spline spaces have been heavily studied; cf., e.g., [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], and references therein.

Of particular interest for applications are spline spaces that possess a basis where every spline is supported only on the star of a vertex. (The star of a vertex is the set of triangles sharing that vertex.) Using such bases in applications leads to sparse linear systems. We call such splines *star-supported*. In [1] they are referred to as *minimally supported*, while in [8] they are called *vertex splines.* It is easy to see that for all $d \geq 1$, the spaces $\mathcal{S}_d^0(\triangle)$ have star-supported bases. In addition, for $r \geq 1$, it is known [9], [10] that the spaces $\mathcal{S}_d^r(\triangle)$ possess bases of star-supported splines for all $d \geq 3r+2$. The following complement to this result is the main result of this paper.

THEOREM 1. *Suppose $r \geq 1$ and $d \leq 3r+1$. Then there are triangulations $\triangle$ for which $\mathcal{S}_d^r(\triangle)$ does not have a star-supported basis.*

Our proof of this theorem is based on showing that there exist triangulations such that the number of linearly independent star-supported splines in the space $\mathcal{S}_d^r(\triangle)$ is less than the dimension of the space. Clearly, it suffices to work with an upper bound on the number of linearly independent star-supported splines and a lower bound on the dimension.

Concerning the dimension of $\mathcal{S}_d^r(\triangle)$, as shown in [13],

$$\begin{aligned}\dim \mathcal{S}_d^r(\triangle) \geq & V_B(d^2+d-2rd+r^2-r)/2 + V_I(d^2-3rd+2r^2) \\ & + 3rd - d^2 - 3r(r-1)/2 + 1 + \sigma,\end{aligned} \tag{1}$$

where V_B and V_I are the number of *boundary vertices* and *interior vertices* of $\triangle$,

[*] Received by the editors July 27, 1998; accepted for publication (in revised form) May 5, 1999; published electronically January 25, 2000.

http://www.siam.org/journals/sima/31-2/34234.html

[†] Department of Mathematics, University of Utah, Salt Lake City, UT 84112 (alfeld@math.utah.edu, http://www.math.utah.edu/~alfeld/).

[‡] Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (s@mars.cas.vanderbilt.edu, http://www.math.vanderbilt.edu/~schumake/). This author was supported by the National Science Foundation under grant DMS-9500643.

respectively, and

$$\sigma = \sum_{v \in \mathcal{V}_I} \sigma_v, \qquad \sigma_v = \sum_{j=1}^{d-r} (r + j + 1 - je_v)_+ . \tag{2}$$

Here $\mathcal{V}_I$ is the set of interior vertices, and e_v is the number of edges of *different slopes* attached to the vertex v.

To help simplify the proof, we shall work with *uniform type*-I *triangulations.* Such a triangulation is obtained by starting with a rectangular grid, which we may assume is generated by the lines $x_i = i/L$ and $y_j = j/L$ for $i, j = 0, \ldots, L$, and then drawing in all diagonals in the northeasterly direction. For a uniform type-I triangulation, the number of edges attached to each interior vertex v is six, and the number of edges with different slopes is three. Thus,

$$\sigma_v = \sum_{j=1}^{d-r} (r + 1 - 2j)_+ \qquad \text{for all } v \in \mathcal{V}_I. \tag{3}$$

Moreover, for this type of triangulation, the number of interior vertices is significantly larger than the number of boundary vertices when L is large, and the term involving V_I dominates in (1). In view of this, to prove Theorem 1 it suffices to establish the following theorem.

Theorem 2. *Let $\triangle_H$ be the triangulation formed by the six triangles surrounding a typical interior vertex v of a type*-I *triangulation. Suppose $r \geq 1$ and $d \leq 3r + 1$, and let*

$$\mathcal{V}_d^r(\triangle_H) = \{s \in \mathcal{S}_d^r(\triangle_H) : s \textit{ vanishes up to order } r \textit{ on the boundary of } \triangle_H\} .$$

Then

$$\dim \mathcal{V}_d^r(\triangle_H) < N_{r,d}(v) := d^2 - 3rd + 2r^2 + \sigma_v. \tag{4}$$

Theorem 2 (and thus also Theorem 1) is trivial in the case $d \leq r$ since in this case $\mathcal{S}_d^r(\triangle) \equiv \mathcal{P}_d$, and clearly there are no star-supported splines in the space. We give a proof of Theorem 2 for $r + 1 \leq d \leq 2r$ in section 2 and for $2r + 1 \leq d \leq 3r + 1$ in section 4.

Throughout the paper we assume familiarity with the Bernstein–Bézier machinery as used, e.g, in [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. In particular, given a vertex v of $\triangle$, we recall that the *jth ring $\mathcal{R}_j(v)$ around v* is the set of domain points at a distance j from v, while the *jth disk $\mathcal{D}_j(v)$ around v* is the set of domain points at a distance of at most j from v. For each domain point P, we write $\lambda_P s$ for the associated coefficient of a spline s. We recall that a subset Γ of the domain points associated with a spline space $\mathcal{S}$ is called a *determining set* for $\mathcal{S}$ provided that the identically zero spline is the only spline $s \in \mathcal{S}$ whose coefficients $\lambda_P s$ are zero for all $P \in \Gamma$. We also recall that if Γ is a determining set, then $\dim \mathcal{S} \leq \#\Gamma$.

We conclude this section with an example to illustrate the basic ideas. Figure 1 shows the B-net of a typical spline in the space$\mathcal{V}_4^1(\triangle_H)$. Coefficients marked with dots on the outermost two rings must be zero because we require function values and first derivatives of a spline in $\mathcal{V}_4^1(\triangle_H)$ to vanish on the boundary of $\triangle_H$. We can identify the points in rings $\mathcal{R}_0(v), \ldots, \mathcal{R}_2(v)$ with the B-net of a spline in $S_2^1(\triangle_H)$. The subset

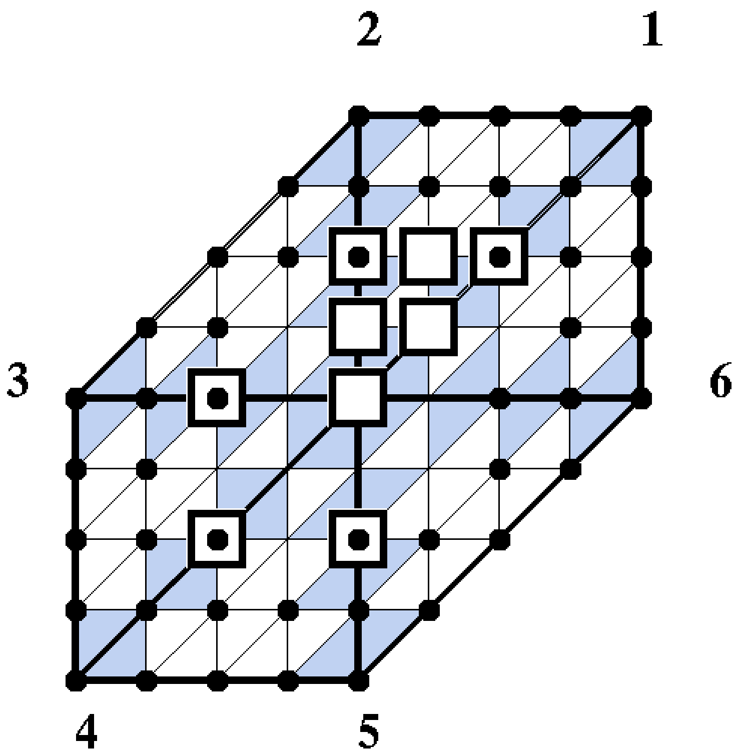


FIG. 1. *A determining set for* $\mathcal{V}_4^1(\triangle_H)$.

of points which are marked with a box in the figure form a minimal determining set for $\mathcal{S}_2^1(\triangle_H)$. This follows from the general theory of minimal determining sets for spline spaces on vertex stars given in [14] but can also easily be verified directly. For a spline $s \in \mathcal{V}_4^1(\triangle_H)$, not all of these coefficients can be set independently, since the smoothness conditions coupled with the boundary conditions imply that certain coefficients in the second ring must be automatically zero. In particular, the C^1 conditions indicated by the quadrilaterals in Figure 1 force the coefficients in the centers of the interior edges to vanish. We see that five coefficients associated with points in the minimal determining set of $\mathcal{S}_2^1(\triangle_H)$ must vanish. These are marked with boxes containing dots. Thus the dimension of $\mathcal{V}_4^1(\triangle_H)$ cannot exceed the number of remaining empty boxes, which is 4. Since $N_{1,4} = 6$, this establishes Theorem 2 in this case.

2. Proof of Theorem 1 for $r + 1 \leq d \leq 2r$. Given an interior vertex v of a triangulation, let $\triangle_v$ be the triangulation consisting of the triangles which make up star(v). Let

$$(5) \qquad \mathcal{V}_d^r(\triangle_v) = \{s \in \mathcal{S}_d^r(\triangle_v) : s \text{ vanishes up to order } r \text{ on the boundary of } \triangle_v\}.$$

In this section we show that for $r + 1 \leq d \leq 2r$, $\mathcal{V}_d^r(\triangle_v)$ is trivial in the sense that it contains only the zero function. Applying this to an interior vertex v of a type-I triangulation $\triangle_I$ and noting that $N_{r,d}$ is positive, this implies Theorem 2 and thus also Theorem 1 in this case. We have the following slightly more precise result.

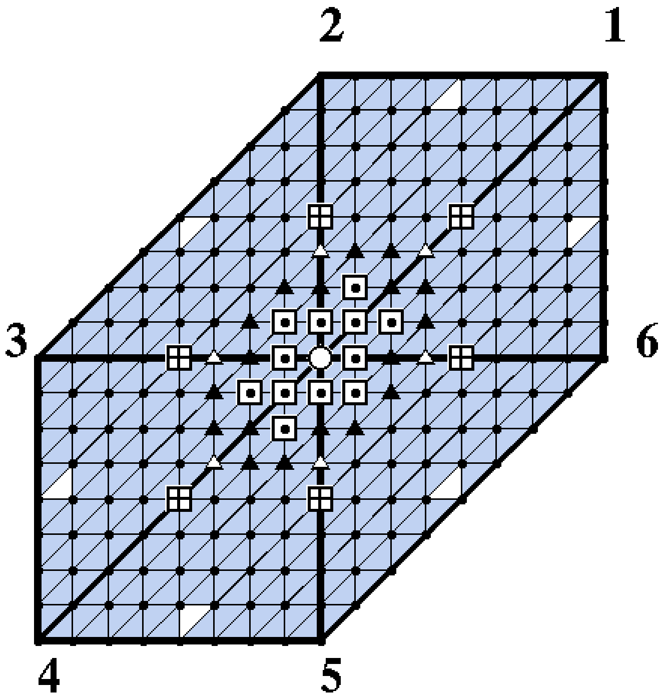


FIG. 2. *Theorem 3 for* $\mathcal{V}_8^4(\triangle_H)$.

THEOREM 3. $\mathcal{V}_d^r(\triangle_v) \equiv \{0\}$ *for all* $d \le 2r$ *if* r *is even and for all* $d \le 2r+1$ *if* r *is odd.*

Proof. Consider first the case $r = 2m$. Suppose $v_1, \ldots, v_n$ are the vertices connected to v, and let $s \in \mathcal{V}_d^r(\triangle_v)$. Then by the boundary conditions, all coefficients of s associated with domain points on the rings $\mathcal{R}_{d-r}(v), \ldots, \mathcal{R}_d(v)$ are zero. Let $w_i^{(1)} = (rv + (d-r)v_i)/d$ for $i = 1, \ldots, n$. Applying Lemma 4 below to the rings $\mathcal{R}_j(w_1^{(1)}), \ldots, \mathcal{R}_j(w_n^{(1)})$ for $j = 1, \ldots, m$ shows that all coefficients of s are zero for domain points on these rings. Then all of the coefficients associated with points on the ring $\mathcal{R}_{d-r-1}(v)$ are zero if and only if $d \le 2r$. Now the process can be repeated based on the points $w_i^{(2)} = ((r+1)v + (d-r-1)v_i$, $i = 1, \ldots, n$. Repeating this process a total of $d - r - 1$ times, we find that all of the coefficients of s are zero.

The case $r = 2m+1$ is similar. In the first step we apply Lemma 4 to the rings $\mathcal{R}_j(w_1^{(1)}), \ldots, \mathcal{R}_j(w_n^{(1)})$ for $j = 1, \ldots, m+1$. Then all coefficients associated with the ring $\mathcal{R}_{d-r+1}(v)$ are zero if and only if $d \le 2r+1$. We then repeat as before. $\square$

Figure 2 illustrates Theorem 3 for $\mathcal{V}_8^4(\triangle_H)$. The points $w_i^{(1)}$ alluded to in the proof are marked with a plus sign in a box. The boundary conditions imply that the coefficients associated with points on $\mathcal{R}_4(v)$ are zero. Then carrying out the first step of the proof, we see that the coefficients associated with the rings $\mathcal{R}_1(w_i^{(1)})$ and $\mathcal{R}_2(w_i^{(1)})$ are zero. These are marked with open triangles and with filled triangles, respectively. In the second step of the proof we get the coefficients in the rings $\mathcal{R}_2(w_i^{(2)})$ to be zero—these are marked as boxes containing a dot. Finally, in the

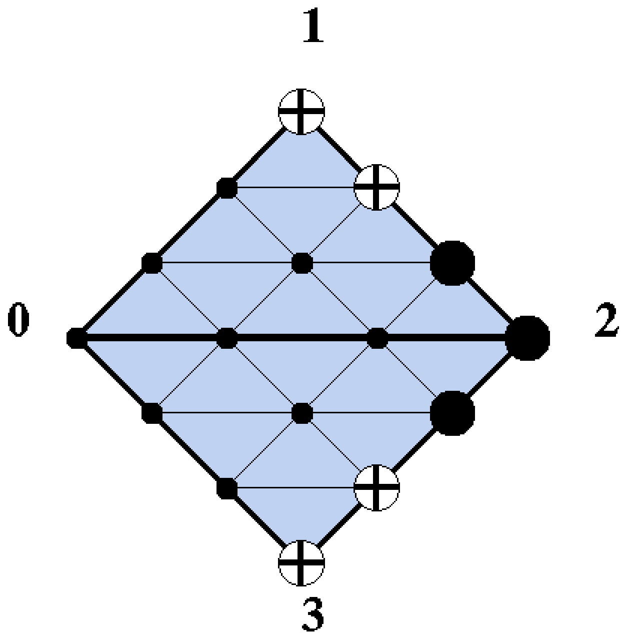


FIG. 3. *Use of Lemma* 4 *for* $j = k = d = 3$ *and* $q = 2$.

third step, we find that the coefficient associated with the point at v (marked with an open circle) is also zero.

The following restatement of Lemma 3.3 of [10] was used in the proof of Theorem 3 and will also be used again later.

LEMMA 4. *Let* $T^{[1]} = \langle v_0, v_1, v_2 \rangle$ *and* $T^{[2]} = \langle v_0, v_2, v_3 \rangle$ *be two triangles sharing the common edge* $e := \langle v_0, v_2 \rangle$. *Suppose* p_1, p_2 *are polynomials of degree* d *on* $T^{[1]}, T^{[2]}$ *which join together with* C^k *smoothness across the edge* e *for some* $0 \le k \le d$. *Given* $k \le j \le d$, *suppose that all coefficients of* p_1 *and* p_2 *in the set* $\mathcal{D}_{j-1}(v_0)$ *are zero, and define*

$$\begin{aligned} c_i &= c^{[1]}_{d-j,i,j-i}, \\ c_{-i} &= c^{[2]}_{d-j,j-i,i}, \end{aligned} \qquad i = 0, \ldots, j.$$

Suppose that

$$c_i = c_{-i} = 0 \quad \text{for } i = k - q + 1, \ldots, k$$

for some q *with* $m = k - 2q \ge -1$. *Suppose in addition that*

$$c_i = 0 \quad \text{for } i = 0, \ldots, m \quad \text{if } m \ge 0.$$

Then $c_i = c_{-i} = 0$ *for all* $i = 0, \ldots, k$.

Figure 3 illustrates Lemma 4 in the case $j = k = d = 3$ and $q = 2$. Here we are assuming that the coefficients associated with the small dots are all zero and that the four coefficients associated with the points marked with a plus sign are also zero. Then the lemma asserts that the three points associated with the large dots must be zero.

3. Constructing minimal determining sets on cells. Let $\triangle_v$ be a triangulation which is obtained by connecting a vertex v to boundary vertices $v_1, \ldots, v_n$. Such a triangulation is called a *cell.* For $\ell = 1, \ldots, n$, let $T^{[\ell]}$ be the triangle with vertices $v, v_\ell, v_{\ell+1}$, where for convenience we identify $v_{n+1} := v_1$. We denote the Bézier points in triangle $T^{[\ell]}$ by $P^{[\ell]}_{ijk}$. We now establish the following modification of Theorem 3.3 in [14].

THEOREM 5. *Let Γ_0 be the set of all Bézier domain points in the triangle $T^{[1]}$. Suppose $\mu_{n-e+1} < \cdots < \mu_n = n+1$ are such that the associated edges are pairwise noncollinear, and let $\mu_1 < \cdots < \mu_{n-e}$ be a complementary set so that $M = \{\mu_1, \ldots, \mu_n\} = \{2, \ldots, n+1\}$. For each $j = 1, \ldots, d-r$, let Γ_j be the first $nj - (r+j+1) + (r+j+1-je)_+$ points in the ordered set*

$$\{P^{[\mu_1]}_{d-j-r,0,j+r}, \ldots, P^{[\mu_1]}_{d-j-r,j-1,r+1}, \ldots, P^{[\mu_n]}_{d-j-r,0,j+r}, \ldots, P^{[\mu_n]}_{d-j-r,j-1,r+1}\}, \tag{6}$$

and let

$$\Gamma = \Gamma_0 \cup \bigcup_{j=1}^{d-r} \Gamma_j.$$

Then the set Γ is a determining set for $\mathcal{S}^r_d(\triangle_v)$.

Proof. This theorem differs from Theorem 3.3 of [14] inasmuch as the points in each group of 6 are written in reverse order. The proof of this version is nearly identical to the original one. Suppose s is a spline in $\mathcal{S}^r_d(\triangle_v)$ such that the coefficients $\lambda_P s$ corresponding to points $P \in \Gamma$ are all zero. We claim that this implies $s \equiv 0$, and thus Γ is a determining set for $\mathcal{S}^r_d(\triangle_v)$. To see this involves examining the rank of a certain block diagonal matrix A; cf. Lemma 2.1 of [14]. Here the submatrix B_j appearing in the proof of Theorem 2.2 in [14] involves different columns in the last block and now corresponds to a Hermite–Birkhoff interpolation problem of the type described in Lemma 6 below. □

LEMMA 6. *Let $\theta_1 < \cdots < \theta_{l+1}$ and let $0 < k < m$ be integers. Then the Hermite–Birkhoff interpolation problem of finding a polynomial p of degree $n := lm + k - 1$ satisfying*

$$\begin{aligned} p^{(j-1)}(\theta_i) &= r_{ij}, \qquad j = 1, \ldots, m, \qquad i = 1, \ldots, l, \\ p^{(m-j)}(\theta_{l+1}) &= r_{l+1,m-j}, \qquad j = 1, \ldots, k, \end{aligned} \tag{7}$$

is poised, i.e., there exists a unique solution for every choice of the data r_{ij}.

Proof. It suffices to show that the homogeneous problem admits only the solution $p \equiv 0$. Suppose p satisfies 7 with 0 data. Then by Rolle's theorem, we conclude that $q := p^{(m-k)}$ has a k-tuple zero at each θ_i, $i = 1, \ldots, l$, $m-k$ zeros in each interval (θ_i, θ_{i+1}), $i = 1, \ldots, l-1$, and an additional k-tuple zero at θ_{l+1}. Thus q has a total of $lk + (l-1)(m-k) + k = n - (m-k) + 1$ zeros, counting multiplicities. But q is a polynomial of degree $n - (m-k)$, and hence q must be identically zero. Integrating $m-k$ times and using the fact that $p^{(j-1)}(\theta_1) = 0$ for $j = 1, \ldots, m$, we conclude that $p \equiv 0$. □

We will apply Theorem 5 to the hexagonal triangulation $\triangle_H$ where $n = 6$ and $e = 3$. In this case we take $M = \{2, 3, 4, 5, 6, 1\}$.

TABLE 1
Number of points in Γ.

Ring	$T^{[1]}$	$T^{[2]}, T^{[3]}, T^{[4]}$	$r = 2m$	$r = 2m + 1$
$\mathcal{R}_0(v)$	1	0	0	0
$\mathcal{R}_1(v)$	2	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{R}_r(v)$	$r + 1$	0	0	0
$\mathcal{R}_{r+1}(v)$	$r + 2$	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{R}_{r+m}(v)$	$r + m + 1$	m	0	0
$\mathcal{R}_{r+m+1}(v)$	$r + m + 2$	$m + 1$	1	0
$\mathcal{R}_{r+m+2}(v)$	$r + m + 3$	$m + 2$	3	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{R}_{2r}(v)$	$2r + 1$	r	$2m - 1$	$2m$

4. Proof of Theorem 1 for $2r + 1 \le d \le 3r + 1$. Throughout this section we assume that

$$d = 2r + k + 1 \tag{8}$$

with $0 \le k \le r$. Suppose v is an interior vertex of a uniform type-I triangulation, and as before, let $\triangle_H$ be the hexagonal triangulation corresponding to star(v). To prove Theorem 2 (and thus also Theorem 1), we need to show that the dimension of $\mathcal{V}_d^r(\triangle_H)$ is bounded by the number $N_{r,d}$ in (4). First, we observe that for this range of d,

$$\sigma_v = \begin{cases} m^2 & \text{if } r = 2m, \\ m^2 + m & \text{if } r = 2m + 1, \end{cases} \tag{9}$$

and thus

$$N_{r,d} = \begin{cases} (k + m + 1)^2, & r = 2m, \\ (k + m + 1)(k + m + 2), & r = 2m + 1. \end{cases} \tag{10}$$

To get an upper bound on $\dim \mathcal{V}_d^r(\triangle_H)$, we proceed as in the example $\mathcal{V}_4^1(\triangle_H)$ discussed in the introduction. We need to find a set Γ which determines $\mathcal{S}_d^r(\triangle_H)$ on $\mathcal{D}_{d-r-1}$ and then examine which of these points can be dropped in view of the interaction of the smoothness conditions with the boundary conditions. Note that $d - r - 1 = r + k$.

To find a set Γ which determines $\mathcal{S}_d^r(\triangle_H)$ on $\mathcal{D}_{d-r-1}$, we identify the domain points of a spline $s \in \mathcal{S}_d^r(\triangle_H)$ lying in $\mathcal{D}_{d-r-1}$ with the domain points of a spline in $\mathcal{S}_{d-r-1}^r(\triangle_H)$ and then apply Theorem 5. We can choose Γ one ring at a time. The rows marked $\mathcal{R}_i(v)$ in Table 1 give the number of points on the rings $\mathcal{R}_0(v), \dots, \mathcal{R}_{r+k}$. For each ring $\mathcal{R}_i(v)$, Γ includes all of the points on that ring in triangle $T^{[1]}$. The number of such points is $i + 1$ and is listed in the second column of the table. In addition, for each $i = r + 1, \dots, r + k$, Γ also includes the last $i - r$ points on $\mathcal{R}_i(v)$ in the triangles $T^{[2]}, T^{[3]}, T^{[4]}$. The numbers of these points are shown in the third column of the table. If $r = 2m$, Γ also includes the last $2(i - m) - 1$ points on $\mathcal{R}_{r+i}(v)$ in the triangles $T^{[5]}$ for $i = m + 1, \dots, k$. These points are shown in the fourth column of the table. Finally, if $r = 2m + 1$, Γ also includes the last $2(i - m - 1)$ points on $\mathcal{R}_{r+i}(v)$ in the triangles $T^{[5]}$ for $i = m + 2, \dots, k$. These points are shown in the fifth column of the table.

We now show how to select a subset $\widetilde{\Gamma}$ of Γ which is a determining set for $\mathcal{V}_d^r(\triangle_H)$. Since the cardinality of $\widetilde{\Gamma}$ is an upper bound on $\dim \mathcal{V}_d^r(\triangle_H)$, our proof of Theorem 2 will be complete if we show that $\#\widetilde{\Gamma} < N_{r,d}$. There are two cases.

Case 1 $(r = 2m)$. For each $i = 1, \dots, m$, let

$$\Gamma_{i,1} := \{\text{the } 2i \text{ domain points in } \mathcal{R}_{r+k-m+i}(v) \cap T^{[1]} \text{ closest to edge } \langle v, v_1\rangle\},$$

and set

$$\Gamma_{i,j} := \{\text{the } 2i \text{ domain points in } \mathcal{R}_{r+k-m+i}(v) \cap T^{[j-1]} \text{ closest to edge } \langle v, v_j\rangle\}$$

for $j = 2, \dots, 6$. Define

$$\widetilde{\Gamma} := \Gamma \setminus \bigcup_{i=1}^{m} \bigcup_{j=1}^{6} \Gamma_{i,j}.$$

Note that the sets $\Gamma_{i,j}$ may contain some points which are not contained in Γ. To see that $\widetilde{\Gamma}$ is a determining set for $\mathcal{V}_d^r(\triangle_H)$, suppose s is a spline in this space with $\lambda_P s = 0$ for all $P \in \widetilde{\Gamma}$. Then since none of the points of Γ in the rings $\mathcal{R}_0(v), \dots, \mathcal{R}_{r+k-m}(v)$ have been removed, all coefficients of s corresponding to points on these rings are zero. Now combining this with the fact that all coefficients of s on the rings $\mathcal{R}_{r+k+1}(v), \dots, \mathcal{R}_d(v)$ are zero, Lemma 4 implies that all coefficients corresponding to the remaining points in $\Gamma \setminus \widetilde{\Gamma}$ are zero. Then $s \equiv 0$, and we have shown that $\widetilde{\Gamma}$ is a determining set.

We now compute the cardinality of $\widetilde{\Gamma}$. The number of points in $\widetilde{\Gamma}$ on the rings $\mathcal{R}_0(v), \dots, \mathcal{R}_{m+k}(v)$ and lying in triangle $T^{[1]}$ is

$$\kappa_1 := \sum_{i=1}^{m+k+1} i = \binom{m+k+2}{2}.$$

The number of points in $\widetilde{\Gamma}$ on rings $\mathcal{R}_{r+1}(v), \dots, \mathcal{R}_{m+k}(v)$ outside of $T^{[1]}$ is given by

$$\kappa_2 := 3 \sum_{i=r+1}^{m+k} (i - r) = \frac{3(k-m+1)(k-m)_+}{2}.$$

We get the factor 3 since such points occur in each of the triangles $T^{[i]}$ for $i = 2, 3, 4$. Now the number of points in $\widetilde{\Gamma}$ lying in triangle $T^{[1]}$ and on the rings $\mathcal{R}_{m+k+1}(v), \dots, \mathcal{R}_{r+k}(v)$ is given by

$$\kappa_3 := \sum_{i=1}^{m} [m+k+i+1-4i]_+ = \sum_{i=1}^{m} [m+k+1-3i]_+ \le \frac{(m+k)(m+k-1)}{6}.$$

Finally, we count the number of points in $\widetilde{\Gamma}$ which lie outside the triangle $T^{[1]}$ and on the rings $\mathcal{R}_{m+k+1}(v), \dots, \mathcal{R}_{r+k}(v)$. There are no such points near the edge $\langle v, v_6\rangle$. Using the values in the fourth column of Table 1, we get

$$\kappa_4 := 3 \sum_{i=1}^{m} [m+k+i-r-2i]_+ = 3 \sum_{i=1}^{m} [k-m-i]_+ = \frac{3(k-m)(k-m-1)_+}{2}.$$

It follows that $n_{r,d} := \kappa_1 + \kappa_2 + \kappa_3 + \kappa_4$ is an upper bound on the cardinality of $\widetilde{\Gamma}$, and

$$n_{r,d} \le \begin{cases} (2k^2 + 4km + 4k + 2m^2 + 4m + 3)/3 & \text{if } 0 \le k \le m, \\ (11k^2 - 14km + 4k + 11m^2 + 4m + 3)/3 & \text{if } m \le k \le 2m. \end{cases}$$

We claim that $n_{r,d} < N_{r,d}$ for all choices of k and m. To see this, note that $\delta(k) := N_{r,d} - n_{r,d}$ is a quadratic polynomial on each of the intervals $[0, m]$ and $[m, 2m]$. Simple calculus shows that both pieces are positive on their domains.

Case 2 ($r = 2m + 1$). This case is very similar to Case 1. For each $i = 0, \dots, m$, let

$$\Gamma_{i,1} := \{\text{the } 2i + 1 \text{ domain points in } \mathcal{R}_{r+k-m+i}(v) \cap T^{[1]} \text{ closest to edge } \langle v, v_1 \rangle\},$$

and set

$$\Gamma_{i,j} := \{\text{the } 2i + 1 \text{ domain points in } \mathcal{R}_{r+k-m+i}(v) \cap T^{[j-1]} \text{ closest to } \langle v, v_j \rangle\}$$

for $j = 2, \dots, 6$. Define

$$\widetilde{\Gamma} := \Gamma \setminus \bigcup_{i=0}^{m} \bigcup_{j=1}^{6} \Gamma_{i,j}.$$

If $s \in \mathcal{V}_d^r(\triangle_H)$ and $\lambda_P s = 0$ for all $P \in \widetilde{\Gamma}$, then all coefficients corresponding to points on the rings $\mathcal{R}_0(v), \dots, \mathcal{R}_{r+k-m-1}(v)$ and $\mathcal{R}_{r+k+1}(v), \dots, \mathcal{R}_d(v)$ are zero. Then Lemma 4 implies that all coefficients corresponding to the remaining points in $\Gamma \setminus \widetilde{\Gamma}$ are zero. Then $s \equiv 0$, and we have shown that $\widetilde{\Gamma}$ is a determining set.

To compute the cardinality of $\widetilde{\Gamma}$, first we note that κ_1 is the same as in Case 1. Now

$$\kappa_2 := 3 \sum_{i=r+1}^{m+k} (i - r) = \frac{3(k - m - 1)(k - m)_+}{2},$$

$$\kappa_3 := \sum_{i=0}^{m} [m + k + i + 2 - 2(2i + 1)]_+ = \sum_{i=0}^{m} [m + k - 3i]_+ \le \frac{(k + m)^2 + 3(k + m) + 2}{6},$$

and

$$\kappa_4 := 3 \sum_{i=0}^{m} [m + k + i - r + 1 - (2i + 1)]_+ = 3 \sum_{i=0}^{m} [k - m - i]_+ = \frac{3(k - m)(k - m - 1)_+}{2}.$$

This leads to

$$n_{r,d} \le \begin{cases} (2k^2 + 4km + 6k + 2m^2 + 6m + 4)/3 & \text{if } 0 \le k \le m, \\ (11k^2 - 14km - 3k + 11m^2 + 15m + 4)/3 & \text{if } m \le k \le 2m + 1. \end{cases}$$

As in the first case, the difference $\delta(k) := N_{r,d} - n_{r,d}$ is a positive quadratic polynomial on each of the intervals $[0, m]$ and $[m, 2m + 1]$. This completes the proof.

Figure 4 illustrates the choice of $\widetilde{\Gamma}$ for the spaces $\mathcal{V}_7^2(\triangle_H)$ and $\mathcal{V}_{10}^3(\triangle_H)$. The boxes represent points in the set Γ, and the boxes containing dots represent points in the set $\Gamma \setminus \widetilde{\Gamma}$. The numbers of linearly independent splines in $\mathcal{V}_7^2(\triangle_H)$ and $\mathcal{V}_{10}^3(\triangle_H)$, respectively, are bounded by the numbers of empty boxes. The numbers are 14 and 26, respectively (see also Table 2 below).

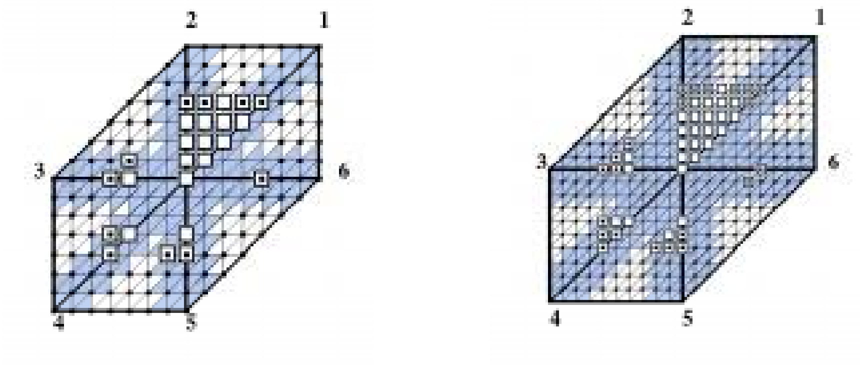


Fig. 4. *Determining sets for* $\mathcal{V}_7^2(\triangle_H)$ *and* $\mathcal{V}_{10}^3(\triangle_H)$.

Table 2
Computed values of D_r, n_r, *and* N_r.

r	D_r	n_r	N_r
1	4	4	6
2	14	14	16
3	25	26	30
4	44	45	49
5	64	66	72
6	92	94	100
7	121	124	132
8	158	161	169
9	196	200	210
10	242	246	256

5. Remarks.

Remark 7. It is known (cf. [5], [6]) that for $d \le 3r+1$ the spline spaces $\mathcal{S}_d^r(\triangle)$ do not have full approximation power on type-I triangulations. This implies that such spaces do not have stable local bases. However, it does not imply the nonexistence of star-supported bases which might be unstable. What we have shown here is that for this range of d, on such triangulations $\mathcal{S}_d^r(\triangle)$ does not possess any basis of star-supported splines, let alone a stable one.

Remark 8. For $d \ge 3r+2$, the spaces $\mathcal{S}_d^r(\triangle)$ have full approximation power, as can be shown by the construction of stable local bases and an associated quasiinterpolant; see [7], [11]. These spaces also have star-supported bases, as was shown in [3]. On the other hand, they do not seem to have stable star-supported bases.

Remark 9. To give an idea of the tightness of our upper bounds on the dimensions of the spaces $\mathcal{V}_d^r(\triangle_H)$, we have used the algebra package REDUCE to compute the dimensions for the case $d = 3r+1$ for $r = 1, \ldots, 10$. The results are displayed in Table 2 which lists the true dimension $D_r := \dim \mathcal{V}_{3r+1}^r(\triangle_H)$, the value of our upper bound on $n_r := n_{r,3r+1}$, and the value of the coefficient $N_r := N_{r,3r+1}$ defined in (4).

Remark 10. As explained in the introduction, to simplify the analysis, we have worked on uniform type-I triangulations $\triangle_I$ generated by $L-1$ interior lines in each direction on a unit square, and we have ignored the number of star-supported splines supported on the stars of boundary vertices. It is not difficult to include such splines in the counts. For example, it is easy to see that for $S_4^1(\triangle_I)$, the total number of star-supported splines is bounded by $4L^2+16L+4$, while the dimension of the space is $6L^2 + 12L + 3$. Thus, we see that $S_4^1(\triangle_I)$ does not admit a star-supported spline basis for any $L \geq 3$. Similarly, for $S_7^2(\triangle_I)$, the total number of star-supported splines is bounded by $14L^2 + 40L + 8$, while the dimension of the space is $16L^2 + 28L + 7$. Thus, we see that $S_7^2(\triangle_I)$ does not admit a star-supported spline basis for any $L \geq 7$.

Remark 11. The problem of computing the exact number of star-supported splines in $\mathcal{S}_d^r(\triangle)$ which are associated with an interior vertex v of Δ is currently under study. The case $r = 1$ is considered in [8].

Remark 12. The figures for this paper were generated using a Java applet which can be found at http://www.math.utah.edu/∼alfeld/MDS/.

REFERENCES

[1] P. Alfeld and L.L. Schumaker, *The dimension of bivariate spline spaces of smoothness r for degree $d \geq 4r + 1$*, Constr. Approx., 3 (1987), pp. 189–197.

[2] P. Alfeld and L.L. Schumaker, *Minimally supported bases for spaces of bivariate piecewise polynomials of smoothness r and degree $d \geq 4r + 1$*, Comput. Aided Geom. Design, 4 (1987), pp. 105–123.

[3] P. Alfeld, B. Piper, and L.L. Schumaker, *An explicit basis for C^1 quartic bivariate splines*, SIAM J. Numer. Anal., 24 (1987), pp. 891–911.

[4] P. Alfeld and L.L. Schumaker, *On the dimension of bivariate splines spaces of smoothness r and degree $d = 3r + 1$*, Numer. Math., 3 (1990), pp. 651–661.

[5] C. de Boor and K. Höllig, *Approximation order from bivariate C^1-cubics: a counterexample*, Proc. Amer. Math. Soc., 87 (1983), pp. 649–655.

[6] C. de Boor and R.Q. Jia, *A sharp upper bound on the approximation order of smooth bivariate pp functions*, J. Approx. Theory, 72 (1993), pp. 24–33.

[7] C.K. Chui, D. Hong, and R.Q. Jia, *Stability of optimal order approximation by bivariate splines over arbitrary triangulations*, Trans. Amer. Math. Soc., 347 (1995), pp. 3301–3318.

[8] C.K. Chui and M.J. Lai, *On bivariate vertex splines*, in Multivariate Approximation Theory III, Internat. Ser. Numer. Math. 73, W. Schempp and K. Zeller, eds., Birkhäuser–Verlag, Basel, Switzerland, 1985, pp. 84–115.

[9] D. Hong, *Spaces of bivariate spline functions over triangulation*, Approx. Theory Appl., 7 (1991), pp. 56–75.

[10] A. Ibrahim and L.L. Schumaker, *Superspline spaces of smoothness r and degree $d \geq 3r + 2$*, Constr. Approx., 7 (1991), pp. 401–423.

[11] M.J. Lai and L.L. Schumaker, *On the approximation power of bivariate splines*, Adv. Comput. Math., 9 (1998), pp. 251–279.

[12] L.L. Schumaker, *On the dimension of multivariate piecewise polynomials*, in Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1986, pp. 1–23.

[13] L.L. Schumaker, *Bounds on the dimension of spaces of multivariate piecewise polynomials*, Rocky Mountain J. Math., 14 (1984), pp. 251–264.

[14] L.L. Schumaker, *Dual bases for spline spaces on cells*, Comput. Aided Geom. Design, 5 (1988), pp. 277–284.

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 467–485

ON THE INITIAL-VALUE PROBLEM IN THE LIFSHITZ–SLYOZOV–WAGNER THEORY OF OSTWALD RIPENING*

BARBARA NIETHAMMER† AND ROBERT L. PEGO‡

Abstract. The Lifshitz–Slyozov–Wagner (LSW) theory of Ostwald ripening concerns the time evolution of the size distribution of a dilute system of particles that evolve by diffusional mass transfer with a common mean field. We prove global existence, uniqueness, and continuous dependence on initial data for measure-valued solutions with compact support in particle size. These results are established with respect to a natural topology on the space of size distributions, one given by the Wasserstein metric which measures the smallest maximum volume change required to rearrange one distribution into another.

Key words. Ostwald ripening, mean-field model, measure-valued solutions, Wasserstein metric

AMS subject classifications. 35L65, 82C21, 82C26, 35Q72, 35D05

PII. S0036141098338211

1. Introduction.

The classical theory of Ostwald ripening, formulated by Lifshitz and Slyozov [6] and Wagner [15], concerns the evolution of the size distribution of a large number of small particles of one phase embedded in a matrix of another phase. Particles are assumed to be widely separated spheres that evolve by diffusional mass transfer with a common mean field. In the late stages of the phase transformation, diffusion is quasi-steady and the particle growth rate is determined by the mass flux at the particle boundary. The mass flux is proportional to the gradient of a potential that is harmonic, is proportional to curvature on the particle boundaries, and is close to constant in the mean field between particles.

In appropriate units, it is found that any particle radius $R(t)$ evolves according to

$$\frac{dR}{dt} = V(R, R_c(t)) := \frac{a}{R^2}\left(\frac{R}{R_c(t)} - 1\right), \tag{1.1}$$

where a is a constant and the critical radius $R_c(t)$ is the same for all particles. The value of $R_c(t)$ is determined from conservation of mass. If mass changes in the diffusion field can be neglected, the particle volume is conserved and one finds that the critical radius equals the average radius of currently existing particles. Particles with radius larger than $R_c(t)$ are growing, and particles with smaller radius shrink and can disappear in finite time.

Classically, the size distribution of particles is described by a particle radius distribution $n(t, R)$. This is a normalized number density that we may scale so that $\int_0^R n(t, r)\, dr$ is the number of (currently existing) particles with radius less than R, divided by the number N of initially existing particles. The number of particles with

[*] Received by the editors May 4, 1998; accepted for publication (in revised form) June 23, 1999; published electronically February 1, 2000. Part of this work was performed at the Max-Planck Institute for Mathematics in the Sciences in Leipzig. This work was partially supported by the NSF under grant DMS-9704924 and the Deutsche Forschungsgemeinschaft through SFB 256.

http://www.siam.org/journals/sima/31-3/33821.html

[†] Institut für Angewandte Mathematik, Universität Bonn, Wegelerstrasse 6, 53115 Bonn, Germany (Barbara.Niethammer@iam.uni-bonn.de).

[‡] Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rpego@ipst.umd.edu).

size between $R_1(t)$ and $R_2(t)$ for any two solutions of (1.1) is conserved, so $n(t,R)$ should satisfy the conservation law

$$\partial_t n + \partial_R(Vn) = 0, \tag{1.2}$$

where the critical radius is given by

$$R_c(t) = \int_0^\infty Rn(t,R)\,dR \bigg/ \int_0^\infty n(t,R)\,dR\,. \tag{1.3}$$

The initial number density $n_0(R) = n(0,R)$ satisfies $\int_0^\infty n_0(R)\,dR = 1$ in this normalization.

A large literature on Ostwald ripening has developed over the past several decades in the metallurgical and physical communities. For an introduction to the problem and an overview of the literature we refer to the review articles of Voorhees [13, 14]. More recently the subject has been taken up by mathematicians. Penrose [10] found that the Lifshitz–Slyozov–Wagner (LSW) evolution law formally governs large-time dynamics in the Becker–Döring clustering equations. The LSW law is derived rigorously in [7] from the Mullins–Sekerka free boundary problem by homogenization methods. The proof suggests that the law is valid only if the electrostatic capacity of the particles is small. In this work the free boundary is restricted to spherical balls, an approach which is justified by a stability analysis in [1]. The works [2, 8, 12] address the large-time asymptotic behavior of the particle size distribution, which is a major point of interest in the theory of Ostwald ripening. The LSW theory predicted asymptotically self-similar behavior, but quantitative agreement with experiment proved elusive.

Our aim in this paper is to develop a satisfactory theory of well-posedness for the initial value problem for the particle size distribution. From the physical point of view, it is reasonable to suppose that a positive fraction of the particles can have the same radius, in which case the size distribution contains one or more Dirac deltas. Mathematically, the ideal is to allow the initial data $n_0(R)\,dR$ to be an arbitrary probability measure such that the total volume $\int_0^\infty \frac{4}{3}\pi R^3 n_0(R)\,dR$ is finite.

It will be convenient to work with particle volume v instead of radius R, and to work with a cumulative number distribution function φ instead of the number density n. We say that

(1.4) φ is the fraction of (initially existing) particles with volume $\geq v$.

As a function of volume v at time t, $\varphi(t,v)$ is a monotonically decreasing function which is left continuous at jumps with $\varphi(t,0) = 1$, and $\int_0^\infty \varphi(t,v)\,dv$ (the total volume) is independent of time. The particle volume distribution, defined by $f(t,v)\,dv = -d\varphi(t,v)$ for each fixed t, is formally related to n via $f(t,v)\,dv = n(t,R)\,dR$.

We normalize the time scale by the factor $4\pi a$ and let $\theta(t) = (4\pi R_c(t)^3/3)^{-1/3}$, so that the volume $v(t)$ of any existing particle should satisfy

$$\frac{dv}{dt} = \Lambda(v,\theta(t)) := v^{1/3}\theta(t) - 1. \tag{1.5}$$

If $v(t)$ is a positive solution of (1.5) on some time interval, then $\varphi(t,v(t))$ should remain constant. This means $\varphi(t,v)$ should be a solution of the hyperbolic equation

$$\partial_t\varphi + \Lambda(v,\theta(t))\partial_v\varphi = 0, \tag{1.6}$$

whose characteristics satisfy (1.5). The value of $\theta(t)$ is obtained from φ in terms of Riemann–Stieltjes integrals by

$$\theta(t) = \int_{0^+}^{\infty} d\varphi(t,v) \Big/ \int_0^{\infty} v^{1/3}\, d\varphi(t,v)\,. \tag{1.7}$$

The numerator is -1 times the quantity $\varphi_0(t) := \lim_{v\to 0}\varphi(t,v)$, which is the fraction of initially existing particles that still exist at time t.

It turns out to be even better to regard the volume v as a function of the fraction φ, $0 \le \varphi \le 1$. We take the map $\varphi \mapsto v(t,\varphi)$ to be right continuous and decreasing with $v(t,1) = 0$. Mathematically, given $\varphi(t,v)$ we obtain $v(t,\varphi)$ via the prescription

$$v(t,x) = \sup\{y \mid \varphi(t,y) > x\} \quad \text{for } 0 \le x < 1 = \max\varphi. \tag{1.8}$$

This is most easily understood when the size distribution corresponds to a finite number of particles. If we list the particle volumes in decreasing order, $v_0(t) \ge \cdots \ge v_{N-1}(t)$, then $v(t,\varphi) = v_j$ for $\varphi \in [j/N,(j+1)/N)$. We shall call $\varphi \mapsto v(t,\varphi)$ a *volume ordering* for the system at time t.

For technical simplicity we shall assume that the particle volumes in the system are bounded. This seems reasonable physically and corresponds to assuming that the particle volume distribution has compact support in v. We then introduce function spaces as follows. Let $rcd([0,1])$ be the set of functions $v : [0,1] \to \mathbb{R}$ that are right continuous, are decreasing, and satisfy $v(1) = 0$. (To be precise, we say v is *decreasing* if $v(x_1) \le v(x_2)$ whenever $x_1 \ge x_2$, and similarly for *increasing*. A decreasing function need not be strictly decreasing.) The set $rcd([0,1])$ is contained in the space $bdd([0,1])$ of real-valued bounded functions on $[0,1]$, equipped with the sup norm $\|v\| = \sup_\varphi |v(\varphi)|$. $rcd([0,1])$ is a complete metric space in the induced topology.

If X is a Banach space and $I \subset \mathbb{R}$ is an interval, then $C(I,X)$ is the space of continuous X-valued functions on I, and $L^\infty_{\text{loc}}(I)$ is the space of equivalence classes of measurable functions locally bounded on I, where two functions are considered equivalent if they agree almost everywhere.

Our main results are the following.

Theorem 1.1 (global existence and uniqueness). *Let $v_0 \in rcd([0,1])$. Then there exist unique functions $\theta \in L^\infty_{\text{loc}}(0,\infty)$ and $v \in C([0,\infty), rcd([0,1]))$, such that*

$$\int_0^1 v(t,\varphi)\, d\varphi = \int_0^1 v_0(\varphi)\, d\varphi$$

for all $t \ge 0$ and

$$v(t,\varphi) = v_0(\varphi) + \int_0^t (v(s,\varphi)^{1/3}\theta(s) - 1)\, ds$$

for all (t,φ) such that $v(t,\varphi) > 0$.

Theorem 1.2 (continuous dependence on initial data). *Given positive constants T and C_0, there exists a positive constant C such that, if (v_1,θ_1) and (v_2,θ_2) are two solutions with the properties stated for (v,θ) in Theorem 1.1, and if $\max(v_1(0,0), v_2(0,0)) \le C_0$, then*

$$\sup_{0\le t\le T} \|v_1(t,\cdot) - v_2(t,\cdot)\| \le C\|v_1(0,\cdot) - v_2(0,\cdot)\|.$$

Consequently the map $v_0 \mapsto v$ is locally Lipschitz from $rcd([0,1])$ into $C([0,\infty), rcd([0,1]))$.

Our strategy to prove these results at the same time justifies a method of numerical approximation for the problem that has a direct physical interpretation. We first consider solutions that are piecewise constant, taking a finite number of values $v_0(t) > \cdots > v_{N-1}(t)$ as is the case for a finite number of particles. We show that these solutions are determined on a succession of time intervals by solving finite systems of coupled ordinary differential equations with a number of components that decreases as the smallest particles vanish. Once we prove the continuity estimate in Theorem 1.2 (at first for initial data near to each other), uniqueness is immediate and existence for general initial data in $rcd([0,1])$ follows by an approximation argument.

The solutions constructed in Theorem 1.1 correspond to measure-valued weak solutions of the evolution equation

$$\partial_t f + \partial_v(\Lambda(v,\theta(t))f) = 0 \tag{1.9}$$

for the particle volume distribution. This means that at each time t, the formal expression $f(t,v)\,dv$ corresponds to a probability measure ν_t having compact support in $[0,\infty)$, the set of volumes. The notion of distance used in Theorem 1.2 has an interpretation as a natural metric on the space $\mathcal{P}_0$ of such probability measures. This metric measures the smallest "maximum volume change" required to rearrange one volume distribution into another. Mathematically it is the L^∞ Wasserstein metric [4, 11], which we denote by d_∞. In section 3 we shall establish the relationship between $v(t,\varphi)$ and ν_t and deduce the following result as a corollary of Theorems 1.1 and 1.2.

THEOREM 1.3. *Let $\mathcal{P}_0$ denote the set of probability measures on $[0,\infty)$ of compact support, with topology given by d_∞, the L^∞ Wasserstein metric. Given $\nu_0 \in \mathcal{P}_0$, there exist a unique $\theta \in L^\infty_{\rm loc}(0,\infty)$ and a unique map $t \mapsto \nu_t$ that is locally Lipschitz from $[0,\infty)$ into $\mathcal{P}_0$, such that (θ,ν) is a volume-conserving weak solution of (1.9), in the sense that*

$$\int_0^\infty v\,d\nu_t(v) = \int_0^\infty v\,d\nu_0(v)$$

for all $t \geq 0$ and

$$\int_0^\infty \int_0^\infty \big(\partial_t\zeta(t,v) + \Lambda(v,\theta(t))\partial_v\zeta(t,v)\big)d\nu_t(v)\,dt = 0$$

for all smooth $\zeta : (0,\infty)\times(0,\infty) \to \mathbb{R}$ with compact support.

Furthermore, given any $T > 0$, $C_0 > 0$, there exists $C > 0$ such that, if two such weak solutions $(\theta_1,\nu^{(1)})$, $(\theta_2,\nu^{(2)})$ are given, such that the supports of $\nu_0^{(1)}$ and $\nu_0^{(2)}$ are contained in $[0,C_0]$, then

$$\sup_{0\leq t\leq T} d_\infty(\nu_t^{(1)},\nu_t^{(2)}) \leq C\,d_\infty(\nu_0^{(1)},\nu_0^{(2)}).$$

It is arguably natural from the physical point of view to measure distance between volume distributions by using the Wasserstein distance as is done here. (In this we were partly motivated by the works [9, 5] that employ Wasserstein distance in connection with parabolic partial differential equations.) A physically reasonable notion of distance should reflect in a plausible way the effect of small perturbations of the system on size distributions. In late-stage Ostwald ripening one imagines that the nucleation or destruction of large particles is unlikely. Thus the topology should not make it "easy" to change the number of large particles. It is plausible, rather, that small perturbations to the system would involve small changes to particle volumes. These notions are captured here by the use of the sup norm distance between volume orderings, and this is equivalent to using the L^∞ Wasserstein metric to compare volume distributions.

In section 4 we briefly treat a related, but simpler, case that arises in LSW theory, in which mass variations in the diffusion field are not neglected. In this case it is not total particle volume that is conserved in time but rather a quantity of the form

$$a\theta(t) + \int_0^1 v(t,\varphi)\,d\varphi,$$

where $a > 0$ is constant. The evolution of particle volumes is still given by (1.5), but θ is now determined directly from the conserved quantity.

2. A priori estimates and well-posedness.

In order to prove the a priori estimate stated in Theorem 1.2, we need a pair of lemmas that yield strengthened variants of Gronwall's inequality.

LEMMA 2.1. *Suppose $G : [0,T] \to \mathbb{R}$ is increasing with $G(0) = 0$, $K \geq 0$ is a constant, and $f : [0,T] \to \mathbb{R}$ is continuous and satisfies*

$$0 \leq f(t) \leq K + \int_{0^+}^t f(s)\,dG(s), \quad 0 \leq t \leq T.$$

Then $f(t) \leq Ke^{G(t)}$ for $0 \leq t \leq T$.

Proof. Let

$$U(t) = K + \int_{0^+}^t f(s)\,dG(s);$$

then $U(0) = K$ and U is increasing. To prove the lemma it suffices to show that $e^{-G}U \leq K$. Let $\{t_j\}_{j=0}^n$ be a partition of $[0,T]$ and define

$$\Delta t = \sup_{1\leq j\leq n}(t_j - t_{j-1}), \quad \epsilon(\Delta t) = \sup_{|t-s|\leq\Delta t}|f(t) - f(s)|.$$

Put $U_j = U(t_j)$, $G_j = G(t_j)$. Then

$$\begin{aligned}
e^{-G_{j+1}}U_{j+1} - e^{-G_j}U_j &= e^{-G_{j+1}}(U_{j+1} - U_j) - U_j(e^{-G_j} - e^{-G_{j+1}})\\
&= e^{-G_{j+1}}\left(\int_{t_j}^{t_{j+1}} f(s)\,dG(s) - U_j\left(e^{G_{j+1}-G_j} - 1\right)\right)\\
&\leq e^{-G_{j+1}}\left(f(t_j) + \epsilon(\Delta t) - U_j\right)(G_{j+1} - G_j)\\
&\leq \epsilon(\Delta t)(G_{j+1} - G_j),
\end{aligned}$$

where we used that $e^x - 1 \geq x$ for all x and $f(t_j) \leq U_j$. Summing, we find that $e^{-G_j}U_j \leq K + \epsilon(\Delta t)G_j$ for all j. Since the partition is arbitrary, $\epsilon(\Delta t)$ can be made arbitrarily small and the result follows. ◻

LEMMA 2.2. *Suppose $G : [0,T] \to \mathbb{R}$ is increasing, and $f : [0,T] \to \mathbb{R}$ is continuous, nonnegative, and increasing. Then as long as $0 \leq t + f(t) \leq T$ we have*

$$\int_0^t (G(s + f(s)) - G(s))\,ds \leq \int_0^{f(0)} (G(f(0)) - G(s))\,ds + \int_0^t f(s)\,d\tilde{G}(s),$$

where $\tilde{G}(s) = G(s + f(s))$.

Proof. Let Q denote the quantity on the right-hand side of the desired inequality. Observe that since G is increasing, we have that

$$Q + \int_t^{t+f(t)} (G(s + f(s)) - G(s))ds \geq Q + \int_t^{t+f(t)} \tilde{G}(s)\, ds - G(t + f(t))f(t).$$

Since $G(t + f(t)) = \tilde{G}(t)$, after integrating by parts in the Riemann–Stieltjes integral and canceling boundary terms we find that the last right-hand side equals

$$\begin{aligned} & -\int_0^{f(0)} G(s)\, ds - \int_0^t \tilde{G}(s)\, df(s) + \int_t^{t+f(t)} \tilde{G}(s)\, ds \\ &= -\int_0^{f(0)} G(s)\, ds - \int_0^t \tilde{G}(s)\, d(s + f(s)) + \int_0^{t+f(t)} \tilde{G}(s)\, ds \\ &= \int_0^{t+f(t)} (G(s + f(s)) - G(s))\ ds. \end{aligned}$$

Canceling the part of the integral from t to $t + f(t)$ finishes the proof. □

Next we establish some basic properties of solutions of the initial value problem as described in Theorem 1.2. Fixing $T > 0$, we shall consider $t \in [0, T]$. Let $\theta \in L^\infty(0, T)$ be positive and let $v \in C([0, T], rcd([0, 1]))$ be such that

$$\int_0^1 v(t, \varphi)\, d\varphi = \int_0^1 v(0, \varphi)\, d\varphi \tag{2.1}$$

for all t and

$$v(t, \varphi) = v(0, \varphi) + \int_0^t (v(s, \varphi)^{1/3}\theta(s) - 1)\, ds \tag{2.2}$$

whenever $v(t, \varphi) > 0$. By scaling, we may assume $\int_0^1 v(t, \varphi)\, d\varphi = 1$ for all $t \in [0, T]$.

From (2.2) it follows that $t \mapsto v(t, \varphi)$ is Lipschitz and satisfies

$$\frac{\partial v}{\partial t} = v^{1/3}\theta(t) - 1 \tag{2.3}$$

for almost every t in any interval where $v > 0$. Since $v^{1/3}\theta - 1 \leq -\frac{1}{2}$ for $v < \varepsilon_0$ where

$$\varepsilon_0^{-1} = 8 \operatorname*{ess\,sup}_{0 \leq t \leq T} \theta(t),$$

it follows easily that if $v(t_0, \varphi) = 0$, then $v(t, \varphi) = 0$ for all $t \geq t_0$.

We define $\bar{v}(t) = v(t, 0) = \max_\varphi v(t, \varphi)$ and with the notation $a \wedge b = \min(a, b)$ we define

$$\begin{aligned} \bar{t}(\varphi) &= \inf\{t \in [0, T] \mid v(t, \varphi) = 0\} \wedge T, \\ \bar{\varphi}(t) &= \sup\{\varphi \in [0, 1] \mid v(t, \varphi) > 0\}. \end{aligned}$$

The functions $\bar{t}$ and $\bar{\varphi}$ are decreasing functions, and $\bar{\varphi}(t) > 0$ for all t, since $v(t, \cdot)$ can never vanish identically by volume conservation. We call $\bar{t}(\varphi)$ the *vanishing time* for $v(t, \varphi)$ at φ if $\bar{t}(\varphi) < T$ (but note that $\bar{t}(\varphi) = T$ if $v(T, \varphi) > 0$).

Lemma 2.3. *For almost every $t \in [0,T]$ we have*

$$0 < \theta(t) = \frac{\bar\varphi(t)}{\displaystyle\int_0^1 v(t,\varphi)^{1/3}\,d\varphi} \le \bar v(t)^{2/3} \le (e^t \bar v(0))^{2/3}.$$

Proof. Evaluate (2.2) at $\min(t,\bar t(\varphi))$ and integrate over $\varphi \in [0,1]$. Changing the order of integration and using the fact that $v(\bar t(\varphi),\varphi) = 0$ if $\bar t(\varphi) < t$, we obtain

$$\begin{aligned} 0 &= \int_0^{\bar\varphi(t)} v(t,\varphi)\,d\varphi - \int_0^1 v(0,\varphi)\,d\varphi \\ &= \int_0^1 \int_0^{\min(t,\bar t(\varphi))} (v(s,\varphi)^{1/3}\theta(s) - 1)\,ds\,d\varphi \\ &= \int_0^t \int_0^{\bar\varphi(s)} (v(s,\varphi)^{1/3}\theta(s) - 1)\,d\varphi\,ds. \end{aligned}$$

Since t is arbitrary the formula for $\theta(t)$ follows. To get the inequalities, we use that $\bar\varphi(t) \le 1$ and $\int_0^1 v^{1/3}\,d\varphi \ge \bar v(t)^{-2/3}\int_0^1 v\,d\varphi$. Then since $d\bar v/dt \le \bar v^{1/3}\theta \le \bar v$ we find

$$\bar v(t) \le e^t \bar v(0). \qquad \square \tag{2.4}$$

Lemma 2.4. *Whenever $v(t_1,\varphi) < \varepsilon_0$, we have $\partial v/\partial t < -\frac{1}{2}$ for almost every $t \in [t_1, \bar t(\varphi)]$ and*

$$\frac{1}{2}(\bar t(\varphi) - t) \le v(t,\varphi) < \varepsilon_0 - \frac{1}{2}(t - t_1)$$

for all $t \in [t_1, \bar t(\varphi)]$.

Proof. $v < \varepsilon_0$ implies $v^{1/3}\theta - 1 < -\frac{1}{2}$ almost everywhere, and the results follow easily. $\square$

Corollary 2.5. *There is a constant $C = C(T, C_0)$ such that*

$$\int_0^{\bar t(\varphi)} v(t,\varphi)^{-2/3}\,dt \le C$$

for all $\varphi \in [0,1]$. Furthermore, the function β given by

$$\beta(t) = \int_0^{\bar\varphi(t)} v(t,\varphi)^{-2/3}\,d\varphi$$

is finite for almost every $t \in [0,T]$ and $\int_0^T \beta(t)\,dt \le C$.

Proof. The first assertion follows directly from the estimates of the preceding lemma. The second follows from Fubini's theorem. $\square$

Our plan now is first to prove a restricted version of Theorem 1.2 for two solutions that are initially close together. This restricted result will suffice to establish the existence and uniqueness theorem, after which the results of Theorem 1.2 without restriction can be proved.

Proposition 2.6. *Given $T > 0$, $C_0 > 0$, there exist $C > 0$ and $\delta > 0$ such that the bound asserted in Theorem 1.2 holds under the additional assumption that*

$$\|v_1(0,\cdot) - v_2(0,\cdot)\| \le \delta.$$

To start the proof of this restricted version of Theorem 1.2, we suppose that $T, C_0 > 0$ are given and put

$$\varepsilon_1 = (8e^T C_0)^{-1}.$$

We suppose that (θ_1, v_1) and $(\theta_2, v_2) \in L^\infty(0,T) \times C([0,T], rcd([0,1]))$ are two solutions of (2.1) and (2.2) such that $\max(v_1(0,0), v_2(0,0)) \le C_0$. We define

$$M(t) = \sup_{0\le s\le t} \|v_1(s,\cdot) - v_2(s,\cdot)\|$$

and assume that $M(0) < \varepsilon_1$.

Lemma 2.7. *There is a constant $C_1 = C_1(T, C_0)$ such that for $0 \le t \le T$ we have*

$$M(t) \le C_1 \left(M(0) + \int_0^t |\theta_1(s) - \theta_2(s)|\, ds \right).$$

Proof. Fix $\varphi \in [0,1]$. We suppose that $\bar t_1(\varphi) \ge \bar t_2(\varphi)$ without loss of generality. For $t \in [0, \bar t_2(\varphi)]$ we may write

$$\begin{aligned} v_1(t,\varphi) - v_2(t,\varphi) = v_1(0,\varphi) - v_2(0,\varphi) &+ \int_0^t v_2(s,\varphi)^{1/3}(\theta_1(s) - \theta_2(s))\, ds \\ &+ \int_0^t \theta_1(s)(v_1(s,\varphi)^{1/3} - v_2(s,\varphi)^{1/3})\, ds. \end{aligned}$$

Using the bounds above for θ_1 and v_2, and the fact that $|a-b| \le |a^3 - b^3|/a^2$ whenever $a, b > 0$, with $C_* = (e^T C_0)^{1/3}$, we obtain the estimate

$$\begin{aligned} |v_1(t,\varphi) - v_2(t,\varphi)| \le |v_1(0,\varphi) - v_2(0,\varphi)| &+ C_* \int_0^t |\theta_1(s) - \theta_2(s)|\, ds \\ &+ C_*^2 \int_0^t v_1(s,\varphi)^{-2/3} |v_1(s,\varphi) - v_2(s,\varphi)|\, ds. \end{aligned} \tag{2.5}$$

For $t \in [\bar t_2(\varphi), \bar t_1(\varphi)]$, we have $v_2(t,\varphi) = 0$ and may write

$$v_1(t,\varphi) \le v_1(\bar t_2(\varphi), \varphi) + C_*^2 \int_{\bar t_2(\varphi)}^t v_1(s,\varphi)^{1/3}\, d\varphi.$$

Using (2.5) with $t = \bar t_2(\varphi)$ to estimate $v_1(\bar t_2(\varphi), \varphi)$, we find that (2.5) is valid for all $t \in [0, \bar t_1(\varphi)]$. Gronwall's inequality then yields that

$$\begin{aligned} &\exp\left(-C_*^2 \int_0^t v_1(s,\varphi)^{-2/3} ds \right) |v_1(t,\varphi) - v_2(t,\varphi)| \\ &\quad \le |v_1(0,\varphi) - v_2(0,\varphi)| + C_* \int_0^t |\theta_1(s) - \theta_2(s)|\, ds. \end{aligned}$$

Using Corollary 2.5 completes the proof. ∎

Lemma 2.8. *Suppose $M(t) \le \varepsilon_1$ for $0 \le t \le \tau$. Then*

$$|\bar\varphi_1(t) - \bar\varphi_2(t)| \le \bar\varphi_1(t) - \bar\varphi_1(t + 2M(t)) + \bar\varphi_2(t) - \bar\varphi_2(t + 2M(t))$$

as long as $t + 2M(t) \le \tau$.

Proof. Fixing t, by relabeling we can assume $\bar\varphi_1(t) \le \bar\varphi_2(t)$. For $\varphi \in [\bar\varphi_1(t), \bar\varphi_2(t)]$, $s \in [t, \tau]$ we have $v_1(s,\varphi) = 0$ and $v_2(s,\varphi) \le M(s) \le \varepsilon_1$ by assumption. By Lemma 2.4, for $s \le \bar t_2(\varphi)$ we have $\partial v_2/\partial t \le -\frac{1}{2}$ and therefore $\bar t_2(\varphi) \le \min(t + 2M(t), T)$. Hence $\bar\varphi_2(t + 2M(t)) \le \bar\varphi_1(t)$, and the result follows. □

LEMMA 2.9. *There is a constant* $C_2 = C_2(T, C_0)$ *and an increasing function* $H : [0,T] \to \mathbb{R}$ *depending on* v_1 *and* v_2*, satisfying* $H(0) = 0$ *and* $H(T) \le C_2$*, such that if* $M(t) \le \varepsilon_1$ *for* $0 \le t \le \tau$*, then*

$$\int_0^t |\theta_1(s) - \theta_2(s)|\, ds \le C_2 M(0) + \int_{0^+}^t M(s)\, dH(s)$$

as long as $t + 2M(t) \le \tau$.

Proof. Using that $\int v_j^{1/3} d\varphi \ge \bar v_j^{-2/3} \ge C_*^{-2}$, from the formula for $\theta(t)$ we obtain that

$$|\theta_1(t) - \theta_2(t)| \le C_*^2 |\bar\varphi_1(t) - \bar\varphi_2(t)| + C_*^4 \int_0^1 |v_1^{1/3} - v_2^{1/3}|\, d\varphi.$$

Let $\varphi_+(t) = \max(\bar\varphi_1(t), \bar\varphi_2(t))$; then for $\varphi < \varphi_+$ we have

$$|v_1^{1/3} - v_2^{1/3}| \le \frac{|v_1 - v_2|}{v_1^{2/3} + v_2^{2/3}}.$$

Note that from Corollary 2.5, it follows that with $t_+(\varphi) = \max(\bar t_1(\varphi), \bar t_2(\varphi))$ we have

$$\int_0^{t_+(\varphi)} \frac{1}{v_1^{2/3} + v_2^{2/3}} dt \le C(T, C_0).$$

By Fubini's theorem it follows that the function defined by

$$h_0(t) = \int_0^{\varphi_+(t)} \frac{1}{v_1^{2/3} + v_2^{2/3}} d\varphi$$

is finite for a.e. t and is integrable with $\int_0^T h_0(t)\, dt \le C(T, C_0)$. Then we have

$$\int_0^1 |v_1^{1/3} - v_2^{1/3}|\, d\varphi \le M(t) h_0(t) \tag{2.6}$$

for a.e. $t \in [0,T]$.

Next, for $j = 1$ and 2 we invoke Lemma 2.2 with $G(t) = -\bar\varphi_j(t)$, $f(t) = 2M(t)$, and conclude that as long as $t + 2M(t) \le \tau$, then

$$\int_0^t \bar\varphi_j(s) - \bar\varphi_j(s + 2M(s))\, ds \le 2M(0) + \int_{0^+}^t 2M(s)\, dH_j(s),$$

where $H_j(t) = -\bar\varphi_j(t + 2M(t)) + \bar\varphi_j(2M(0))$. Evidently H_j satisfies $H_j(t) \le 1$ for all t.

Putting these estimates together with the result of Lemma 2.8, we find that

$$\int_0^t |\theta_1(s) - \theta_2(s)|\, ds \le 4C_*^2 M(0) + \int_{0^+}^t M(s)\, dH(s),$$

where

$$H(t) = 2C_*^2(H_1(t) + H_2(t)) + C_*^4 \int_0^t h_0(s)\, ds.$$

The desired result follows. □

The proof of Proposition 2.6 uses a continuation argument based on the estimates above together with the estimate

$$M(\tau) - M(t) \le 2C_*^3(\tau - t) \tag{2.7}$$

whenever $0 \le t \le \tau \le T$, which follows from $|\partial v/\partial t| \le C_*^3$. Since M is increasing, we can find $\tilde{T} \le T$ such that $\tilde{T} + 2M(\tilde{T}) = T$. With $\tau = t + 2M(t)$, inequality (2.7) yields

$$M(t + 2M(t)) \le M(t)(1 + 4C_*^3) \tag{2.8}$$

whenever $t \le \tilde{T}$. Now let

$$\Omega = \{t \in [0, \tilde{T}] \mid M(t + 2M(t)) \le \varepsilon_1\}.$$

If $M(0) \le \delta_0 := \varepsilon_1/(1 + 4C_*^3)$, then $0 \in \Omega$ so Ω is nonempty, and clearly Ω is closed. We claim Ω is open in $[0, \tilde{T}]$ if $M(0)$ is sufficiently small.

Given any $t_1 \in \Omega$ we can apply Lemmas 2.7 and 2.9 to deduce that

$$M(t) \le C_1(1 + C_2)M(0) + C_1 \int_{0^+}^t M(s)\, dH(s) \tag{2.9}$$

for $0 \le t \le t_1$. Then Lemma 2.1 implies

$$M(t) \le C_3 M(0) \tag{2.10}$$

for $0 \le t \le t_1$, where $C_3(T, C_0) = \exp(C_1C_2)C_1(1 + C_2)$. Using (2.8) we infer that $M(t_1 + 2M(t_1)) \le C_4 M(0)$ with $C_4 = C_3(1 + 4C_*^3)$. Provided we assume

$$M(0) \le \delta_1 := \frac{1}{2}\frac{\varepsilon_1}{C_4},$$

it follows that $M(t_1 + 2M(t_1)) < \varepsilon_1$, and since M is continuous, Ω is open in $[0, \tilde{T}]$.

Consequently we have $\tilde{T} \in \Omega$. Putting $t_1 = \tilde{T}$, this means we have $M(T) \le \varepsilon_1$ and $M(T) \le C_4 M(0)$ if $M(0) \le \delta_1$. This finishes the proof of Proposition 2.6.

Proof of Theorem 1.1. Uniqueness follows immediately from Proposition 2.6. To prove existence for arbitrary $v_0 \in rcd([0,1])$, by Proposition 2.6 and Lemma 2.3 it evidently suffices to prove global existence for v_0 in a dense set of $rcd([0,1])$. Solutions in general are constructed by passing to the limit in $C([0,T], rcd([0,1]))$ for every $T > 0$.

Lemma 2.10. *The set of functions in* $rcd([0,1])$ *that take a finite number of values is dense in* $rcd([0,1])$.

Proof. Let $v_0 \in rcd([0,1])$ and let $\varepsilon > 0$. Let $y_j = \frac{1}{2}\varepsilon j$ for $j = 0, 1, \ldots,$ and let

$$v_\varepsilon(\varphi) = \min\{y_j \mid y_j \ge v_0(\varphi)\}$$

for $\varphi \in [0,1]$. It is easy to see that v_ε has a finite number of values, that $v_\varepsilon \in rcd([0,1])$, and $\|v_\varepsilon - v_0\| < \varepsilon$. This proves the lemma. □

Suppose, then, that $v_0 \in rcd([0,1])$ takes a finite number of values $y_0 > \cdots > y_N = 0$. Then with $\varphi_j = \inf\{\varphi \mid v_0(\varphi) = y_j\}$, we have $0 = \varphi_0 < \cdots < \varphi_N \leq 1$ and $v_0(\varphi) = y_j$ for $\varphi \in [\varphi_j, \varphi_{j+1})$, $j = 0, \dots, N-1$. We start to construct a solution by solving the system of ordinary differential equations

$$w_j'(t) = w_j(t)^{1/3}\Theta(t) - 1, \qquad j = 0, \dots, N-1, \tag{2.11}$$

with

$$\Theta(t) = \varphi_N \Bigg/ \sum_{j=0}^{N-1} w_j(t)^{1/3}(\varphi_{j+1} - \varphi_j) \tag{2.12}$$

and $w_j(0) = y_j$, on a maximal interval $[0, t_N)$ in which $\min w_j(t) > 0$. The solution is smooth and $w_j(t) > w_{j+1}(t)$ by backward uniqueness for the equation $w' = w^{1/3}\Theta - 1$. The quantity

$$\sum_{j=0}^{N-1} w_j(t)(\varphi_{j+1} - \varphi_j)$$

is conserved in time. Without loss of generality we can assume this quantity is 1.

We can estimate $\Theta(t) \leq w_0(t)^{2/3}$ so $w_0' \leq w_0$ and hence $w_0(t) \leq e^t y_0$. If $t_N < \infty$, then it follows that the smallest component vanishes, i.e., $w_{N-1}(t_N^-) = \lim_{t \nearrow t_N} w_{N-1}(t) = 0$.

For $t \in [0, t_N)$ we define $v(t, \varphi) = w_j(t)$ for $\varphi \in [\varphi_j, \varphi_{j+1})$, $j = 0, \dots, N-1$, and let $\theta = \Theta$. This yields a solution of (2.2) and (2.1) for $t \in [0, t_N)$. As $t \to t_N$ from below, the limits $v(t_N^-, \varphi)$ and $\theta(t_N^-)$ exist. The solution can then be reinitialized at time t_N with one less component (N replaced by $N-1$). After some finite number of such steps the solution must exist globally.

Thus, for $v_0 \in rcd([0,1])$ with a finite number of values, a global solution exists. Theorem 1.1 follows.

Proof of Theorem 1.2. The additional restriction imposed in Proposition 2.6 can be removed now by considering convex combinations of initial data. Given T, C_0, v_1, and v_2 as stated, let $C > 0$, $\delta > 0$ be as given by Proposition 2.6, and let $M_0 = \|v_1(0,\cdot) - v_2(0,\cdot)\|$. Fix an integer $n > M_0/\delta$, and for $j = 0, 1, \dots, n$ let

$$x_j(\varphi) = \left(1 - \frac{j}{n}\right) v_1(0,\varphi) + \left(\frac{j}{n}\right) v_2(0,\varphi)$$

for $\varphi \in [0,1]$. Then $x_j \in rcd([0,1])$, $x_j(0) \leq C_0$ for all j and $\|x_{j+1} - x_j\| = M_0/n < \delta$. By the existence theorem there exist corresponding solutions $v = \tilde{v}_j$ to (2.1)–(2.2) with $\tilde{v}_j(0,\cdot) = x_j$, and by Proposition 2.6 we have

$$\sup_{0 \leq t \leq T} \|\tilde{v}_{j+1}(t,\cdot) - \tilde{v}_j(t,\cdot)\| \leq C\|x_{j+1} - x_j\| = CM_0/n.$$

Since $v_1 - v_2 = \sum_{j=0}^{n-1}(\tilde{v}_{j+1} - \tilde{v}_j)$, using the triangle inequality we find that

$$\sup_{0 \leq t \leq T} \|v_1(t,\cdot) - v_2(t,\cdot)\| \leq CM_0,$$

as desired.

3. Measure-valued solutions. Our aim here is to describe a precise correspondence between the solutions $v(t,\varphi)$ of Theorem 1.1 and measure-valued weak solutions ν_t of (1.9) and to show that the metric $\|v_1 - v_2\|$ on $rcd([0,1])$ corresponds to the L^∞ Wasserstein metric on the space $\mathcal{P}_0$ of (Borel) probability measures on $[0,\infty)$ with compact support. Theorem 1.3 then follows as a corollary of Theorems 1.1 and 1.2.

We begin with a technical lemma on generalized inverses of increasing functions.

Lemma 3.1. *Suppose $b > 0$ and $w : [0,b] \to \mathbb{R}$ is a left continuous increasing function with $w(0) = 0$. Let $b^\dagger = w(b)$ and define $w^\dagger : [0,b^\dagger] \to \mathbb{R}$ by*

$$w^\dagger(y) = \begin{cases} \sup\{x \mid w(x) < y\}, & 0 < y \le b^\dagger, \\ 0, & y = 0. \end{cases}$$

Then $w^\dagger$ is left continuous and increasing and, moreover,

$$w^{\dagger\dagger} = w.$$

Proof. Clearly $w^\dagger$ is increasing. Given $y \in (0,b^\dagger]$ and $\varepsilon > 0$, put $\bar{x} = w^\dagger(y)$ and $2\delta = y - w(\bar{x} - \varepsilon)$. Then $\delta > 0$ and $w(\bar{x} - \varepsilon) < y - \delta$; hence $\bar{x} - \varepsilon < w^\dagger(y - \delta) \le \bar{x}$. It follows $w^\dagger$ is left continuous.

To show $w^{\dagger\dagger} = w$, it suffices to show that for $0 < x < b$,

$$w(x - \varepsilon) \le w^{\dagger\dagger}(x) \le w(x + \varepsilon)$$

for all sufficiently small $\varepsilon > 0$. Let $\bar{y} = w^{\dagger\dagger}(x) = \sup\{y \mid w^\dagger(y) < x\}$. Then for all $\varepsilon_0 > 0$, $w^\dagger(\bar{y}+\varepsilon_0) \ge x$; hence for any small $\varepsilon > 0$ we have $w(x-\varepsilon) < \bar{y}+\varepsilon_0$. Therefore $w(x - \varepsilon) \le \bar{y}$.

For the reverse inequality there are two cases. If $\bar{x} = w^\dagger(\bar{y}) < x$, then for small $\varepsilon > 0$ we have

$$\bar{y} \le w(\bar{x} + \varepsilon) \le w(x + \varepsilon). \tag{3.1}$$

Otherwise $\bar{x} \ge x$, and since w is left continuous, $\bar{x} = x$. In this case, (3.1) again holds. This finishes the proof. □

If w is continuous and strictly increasing, then $w^\dagger$ is the inverse function of w.

Given a probability measure ν with compact support $[0,\bar{v}] \subset [0,\infty)$, we associate the distribution function $F_\nu : [0,\infty) \to [0,1]$ given by

$$F_\nu(x) = \begin{cases} \nu([0,x)), & x > 0, \\ 0, & x = 0. \end{cases} \tag{3.2}$$

F_ν is left continuous and increasing, and F_ν determines ν (that is, the values of F_ν determine the values of ν on all Borel sets). We associate a decreasing function $v = \hat{v}(\nu)$ to ν via $v(x) = F_\nu^\dagger(1-x)$ for $x \in [0,1]$. (Here, $F_\nu^\dagger$ is the generalized inverse of the restriction of F_ν to $[0,\bar{v}+1]$.) That is,

$$v(x) = \begin{cases} \sup\{y \mid F_\nu(y) < 1 - x\}, & 0 \le x < 1, \\ 0, & x = 1. \end{cases} \tag{3.3}$$

With the notation $Rv(x) = v(1-x)$ we have $\hat{v}(\nu) = R(F_\nu^\dagger)$. The first part of Lemma 3.1 implies $\hat{v}(\nu) \in rcd([0,1])$; thus the map $\hat{v} : \mathcal{P}_0 \to rcd([0,1])$. (Recall $\mathcal{P}_0$ is the set of probability measures on $[0,\infty)$ with compact support.)

The inverse map to $\hat{v}$ is given as follows. If $v \in rcd([0,1])$ we let $F = (Rv)^\dagger$ on $[0, v(0)]$ and put $F(x) = 1$ for $x > v(0)$. Then F is increasing and left continuous and determines a (Borel) probability measure ν for which $F = F_\nu$. For later use we note that for any continuous $f : (0,\infty) \to \mathbb{R}$ with compact support, we have

$$\int_0^1 f(v(x))\,dx = \int_0^1 f(F^\dagger(x))\,dx = \int_0^\infty f(y)\,dF(y) = \int_0^\infty f(y)\,d\nu(y). \tag{3.4}$$

This follows from [3, 2.5.18(3)], for example. The identity function $y \mapsto y$ can be approximated uniformly on compact sets in $[0,\infty)$ by such functions f, hence

$$\int_0^1 v(x)\,dx = \int_0^\infty y\,d\nu(y). \tag{3.5}$$

We let $\hat{\nu}(v) = \nu$, so $\hat{\nu} : rcd([0,1]) \to \mathcal{P}_0$. Lemma 3.1 implies that we have the following.

Lemma 3.2. *$\hat{v}$ and $\hat{\nu}$ are inverse maps: $\hat{v}(\hat{\nu}(v)) = v$ for all $v \in rcd([0,1])$ and $\hat{\nu}(\hat{v}(\nu)) = \nu$ for all $\nu \in \mathcal{P}_0$.*

We now recall from [4] that the L^p Wasserstein metric can be defined on $\mathcal{P}_0$ as follows. Given ν_1 and ν_2 in $\mathcal{P}_0$, let $D(\nu_1,\nu_2)$ be the set of probability measures μ on $[0,\infty) \times [0,\infty)$ with marginal distributions ν_1 and ν_2, that is, for all continuous $\zeta : [0,\infty) \to \mathbb{R}$,

$$\int_0^\infty \int_0^\infty \zeta(x)\,d\mu(x,y) = \int_0^\infty \zeta(x)\,d\nu_1(x)$$

and

$$\int_0^\infty \int_0^\infty \zeta(y)\,d\mu(x,y) = \int_0^\infty \zeta(y)\,d\nu_2(y).$$

If $1 \le p < \infty$, then the L^p Wasserstein metric is defined by

$$d_p(\nu_1,\nu_2) = \left(\inf_{\mu \in D(\nu_1,\nu_2)} \int |x-y|^p\,d\mu(x,y) \right)^{1/p}.$$

The L^∞ Wasserstein metric is defined by

$$d_\infty(\nu_1,\nu_2) = \inf_{\mu \in D(\nu_1,\nu_2)} \mu\text{-ess sup}|x-y|.$$

The measures μ represent ways to "rearrange mass" from one distribution into the other, and the L^p Wasserstein metrics measure the least costly way to do this according to the notion of cost indicated.

Lemma 3.3. *Given ν_1, $\nu_2 \in \mathcal{P}_0$, let $v_1 = \hat{v}(\nu_1)$, $v_2 = \hat{v}(\nu_2)$. Then for $1 \le p < \infty$ we have*

$$d_p(\nu_1,\nu_2) = \left(\int_0^1 |v_1(\varphi) - v_2(\varphi)|^p\,d\varphi \right)^{1/p},$$

and

$$d_\infty(\nu_1,\nu_2) = \|v_1 - v_2\|.$$

Proof. The assertion for $1 \le p < \infty$ follows from [11, pp. 28–30, Corollary 7.3.6], which yields that

$$d_p(\nu_1, \nu_2) = \left(\int_0^1 |F^\dagger_{\nu_1}(\varphi) - F^\dagger_{\nu_2}(\varphi)|^p \, d\varphi \right)^{1/p}.$$

Then Proposition 3 of [4] asserts that $\lim_{p\to\infty} d_p(\nu_1, \nu_2) = d_\infty(\nu_1, \nu_2)$. Since v_1 and v_2 are right continuous, it follows

$$\begin{aligned} d_\infty(\nu_1, \nu_2) &= \lim_{p\to\infty} \left(\int_0^1 |v_1(\varphi) - v_2(\varphi)|^p \, d\varphi \right)^{1/p} \\ &= \operatorname*{ess\,sup}_{[0,1]} |v_1(\varphi) - v_2(\varphi)| = \sup_{[0,1]} |v_1(\varphi) - v_2(\varphi)|. \qquad \square \end{aligned}$$

Corollary 3.4. *Let $\mathcal{P}_0$ have the topology induced by d_∞. Then $\mathcal{P}_0$ is complete, and the map $\hat{v} : \mathcal{P}_0 \to rcd([0, 1])$ is an isometric isomorphism of complete metric spaces.*

The completeness of $\mathcal{P}_0$ with respect to the metric d_∞ was established in [4].

The correspondence between volume orderings $v \in rcd([0, 1])$ and volume distributions $\nu \in \mathcal{P}_0$ has been established. Now we seek to show that this correspondence maps solutions to weak solutions and vice versa.

Proposition 3.5. *Let $\theta \in L^\infty_{\text{loc}}(0, \infty)$, $v \in C([0, \infty),\ rcd([0, 1]))$ be a solution of* (2.2)*. For each $t \ge 0$, let $\nu_t = \hat{\nu}(v(t, \cdot))$. Then $\nu : [0, \infty) \to \mathcal{P}_0$ is locally Lipschitz, and ν is a weak solution of* (1.9)*.*

Proof. From (2.2) we have that $v : [0, \infty) \to rcd([0, 1])$ is locally Lipschitz; therefore $\nu : [0, \infty) \to \mathcal{P}_0$ is locally Lipschitz by Corollary 3.4.

Let $\zeta : (0, \infty) \times (0, \infty) \to \mathbb{R}$ be smooth with compact support. Then for all $\varphi \in [0, 1]$, $t \mapsto \zeta(t, v(t, \varphi))$ is Lipschitz continuous and we have

$$\begin{aligned} 0 &= \int_0^\infty \frac{d}{dt} \zeta(t, v(t, \varphi)) \, dt \\ &= \int_0^\infty \partial_t \zeta(t, v(t, \varphi)) + \Lambda(v(t, \varphi), \theta(t)) \partial_v \zeta(t, v(t, \varphi)) \, dt. \end{aligned}$$

Let $F(t, \cdot) = (Rv(t, \cdot))^\dagger$ and let $\varphi(t, \cdot) = 1 - F(t, \cdot)$. We integrate over $\varphi \in [0, 1]$, use Fubini's theorem, and change variables using (3.4). We obtain

$$\begin{aligned} 0 &= -\int_0^\infty \int_0^\infty \left(\partial_t \zeta(t, v) + \Lambda(v, \theta(t)) \partial_v \zeta(t, v) \right) \, d\varphi(t, v) \, dt \\ &= \int_0^\infty \int_0^\infty \left(\partial_t \zeta(t, v) + \Lambda(v, \theta(t)) \partial_v \zeta(t, v) \right) \, d\nu_t(v) \, dt. \end{aligned}$$

Thus (θ, ν) is a weak solution in the sense of Theorem 1.3, as claimed. $\square$

Proposition 3.6. *Suppose that $\theta \in L^\infty_{\text{loc}}(0, \infty)$ and $\nu : [0, \infty) \to \mathcal{P}_0$ is locally Lipschitz, and ν is a weak solution of* (1.9)*. Let $v(t, \cdot) = \hat{v}(\nu_t)$ for each $t \ge 0$. Then v is a solution of* (2.2)*.*

Proof. Given θ and v as described, the map $v : [0, \infty) \to rcd([0, 1])$ is locally Lipschitz. We consider test functions ζ of the form

$$\zeta(t, v) = \xi(t) \eta(v), \tag{3.6}$$

where the functions ξ, $\eta : \mathbb{R} \to \mathbb{R}$ are smooth with compact support in $(0, \infty)$. Using this form together with the fact that (θ, ν) form a weak solution to (1.9), and using the change of variables from (3.4) as previously, we find that

$$0 = \int_0^\infty \int_0^1 \xi'(t)\eta(v(t,\varphi)) + \xi(t)\eta'(v(t,\varphi))\Lambda(v(t,\varphi),\theta(t))\, d\varphi\, dt.$$

Using Fubini's theorem and integrating by parts in time, this gives

$$0 = \int_0^1 \int_0^\infty \xi(t)\tilde{\eta}(v(t,\varphi))(\Lambda(v(t,\varphi),\theta(t)) - \partial_t v(t,\varphi))\, dt\, d\varphi, \tag{3.7}$$

where $\tilde{\eta} = \eta'$. This formula is justified since for each $\varphi \in [0, 1]$, $v(\cdot, \varphi)$ is Lipschitz, hence differentiable almost everywhere. We note that since v is bounded on compact sets, $\tilde{\eta}$ can be chosen to agree on the range of v with an arbitrary smooth function with compact support in $(0, \infty)$. We do this and drop the tilde. Furthermore, we note that by Lebesgue's dominated convergence theorem, (3.7) remains valid for any ξ with compact support in $(0, \infty)$ that is the bounded pointwise limit $\xi = \lim_{n\to\infty} \xi_n$ of a sequence of smooth ξ_n with compact support in $(0, \infty)$, and similarly for η. For the moment it will suffice to consider ξ, $\eta \in C_c(\mathbb{R}^+)$, the set of continuous functions on $(0, \infty)$ with compact support.

For what follows, we take some care regarding joint measurability in (t, φ) and sets of measure zero. We fix a representative $\tilde{\theta}$ in the equivalence class θ, then drop the tilde. $\partial_t v(t, \varphi)$ need not exist at every point, but (3.7) also holds if $\partial_t v$ is replaced by the upper derivative $\overline{\partial}_t v$ or the lower derivative $\underline{\partial}_t v$, defined by

$$\overline{\partial}_t v(t,\varphi) = \lim_{\varepsilon\to 0} \sup_{0<|h|<\varepsilon} \delta^h v(t,\varphi), \quad \underline{\partial}_t v(t,\varphi) = \lim_{\varepsilon\to 0} \inf_{0<|h|<\varepsilon} \delta^h v(t,\varphi),$$

where

$$\delta^h v(t,\varphi) = \frac{v(t+h,\varphi) - v(t,\varphi)}{h}.$$

Since v is locally Lipschitz in t uniformly in φ, $\overline{\partial}_t v$ and $\underline{\partial}_t v$ are bounded on compact sets, and $\underline{\partial}_t v \le \overline{\partial}_t v$.

LEMMA 3.7. *As maps from* $(0, \infty) \times [0, 1] \to \mathbb{R}$, $\overline{\partial}_t v$ *and* $\underline{\partial}_t v$ *are Borel measurable. Moreover,* $\overline{\partial}_t v = \underline{\partial}_t v$ *almost everywhere in* $(0, \infty) \times [0, 1]$.

Proof. Since v is continuous in t uniformly in φ and is right continuous and decreasing in φ, v is lower semicontinuous, hence Borel. Suppose $0 < t_1 < t_2 < \infty$; then for $0 < |h| < t_1$ the map $\delta^h v$ is Borel on $[t_1, t_2] \times [0, 1]$. Let $\{h_j\}$ be a dense sequence in $(-1, 0) \cup (0, 1)$. Since the maximum of two Borel functions is Borel and pointwise limits of sequences of Borel functions are Borel, and pointwise we have

$$\sup_{0<|h|<\varepsilon} \delta^h v(t,\varphi) = \sup_{|h_j|<\varepsilon} \delta^{h_j} v(t,\varphi) = \lim_{k\to\infty} \max_{\substack{j\le k\\ |h_j|<\varepsilon}} \delta^{h_j} v(t,\varphi),$$

by taking ε to zero along a sequence it follows that $\overline{\partial}_t v$ is Borel on $[t_1, t_2] \times [0, 1]$, hence on $(0, \infty) \times [0, 1]$. A similar argument applies for $\underline{\partial}_t v$.

Now we have that the set $Z = \{(t, \varphi) \mid (\overline{\partial}_t v - \underline{\partial}_t v)(t, \varphi) > 0\}$ is a Borel set. We know that for each φ, $v(\cdot, \varphi)$ is differentiable almost everywhere, so $(\overline{\partial}_t v - \underline{\partial}_t v)(t, \varphi) = 0$ for almost every $t > 0$. Fubini's theorem now implies that Z has Lebesgue measure zero in $(0, \infty) \times [0, 1]$. □

Returning to (3.7), we can now apply Fubini's theorem and deduce that for almost every t, $(\overline{\partial}_t v - \underline{\partial}_t v)(t,\varphi) = 0$ for almost every φ, and with

$$J_\eta(t) = \int_0^1 \eta(v(t,\varphi))(\Lambda(v(t,\varphi),\theta(t)) - \overline{\partial}_t v(t,\varphi))\,d\varphi,$$

we have that for any $\eta \in C_c(\mathbb{R}^+)$, $\int_0^\infty \xi(t)J_\eta(t)\,dt = 0$ for all $\xi \in C_c(\mathbb{R}^+)$. Therefore, given η there is a set $\Omega_\eta \subset (0,\infty)$ of full measure (meaning the complement has measure zero), such that $J_\eta(t) = 0$ for all $t \in \Omega_\eta$.

The set $C_c(\mathbb{R}^+)$ is separable, so if we do this for a dense sequence $\{\eta_n\}$ we find there is a set $\Omega \subset \cap\Omega_{\eta_n}$ of full measure in $(0,\infty)$ such that for $t \in \Omega$, $J_{\eta_n}(t) = 0$ for all n. Since any $\eta \in C_c(\mathbb{R}^+)$ can be approximated uniformly by a subsequence of $\{\eta_n\}$, we infer the following.

Lemma 3.8. *There is a set $\Omega \subset (0,\infty)$ of full measure, such that for all $t \in \Omega$, $\partial_t v(t,\varphi)$ exists for almost every $\varphi \in [0,1]$ and $J_\eta(t) = 0$ for all $\eta \in C_c(\mathbb{R}^+)$.*

Lemma 3.9. *Let $t \in \Omega$, and suppose $v(t,x) = v(t,y)$, where $0 \le x < y \le 1$. Then $\partial_t v(t,\varphi)$ exists for all $\varphi \in (x,y)$ and is constant on this interval.*

The proof of this lemma is straightforward, using the facts that $v(t,\cdot)$ is decreasing for every t and $\partial_t v(t,\varphi)$ exists for almost every φ.

Lemma 3.10. *Let $t \in \Omega$, and let $\bar\varphi(t) = \sup\{\varphi \mid v(t,\varphi) > 0\}$. Then*

$$\Lambda(v(t,\varphi),\theta(t)) - \partial_t v(t,\varphi) = 0$$

for almost all $\varphi \in [0,\bar\varphi(t))$.

Proof. We thank B. Kirchheim for the main idea of the following proof. Since t is fixed, we suppress indicating dependence on t, and we let $g(\varphi) = \Lambda(v(t,\varphi),\theta(t)) - \overline{\partial}_t v(t,\varphi)$. We know g is measurable and bounded. $\varphi \mapsto v(\varphi)$ is decreasing, so if $0 \le y$ is in the range of v, either the preimage $v^{-1}(y)$ is a singleton or an interval of nonzero width. There can be only a countable set of y of the latter type. Let Δ be the (countable) set of endpoints of such intervals. For x, $y \in [0,1] \setminus \Delta$, we know that $v(x) = v(y)$ implies $\overline{\partial}_t v(x) = \overline{\partial}_t v(y)$, and so $g(x) = g(y)$.

Given any $\varepsilon > 0$, let $A_\varepsilon = [0,\bar\varphi(t)) \cap \{x \in [0,1] \mid g(x) > \varepsilon\} \setminus \Delta$. Then A_ε is measurable, and we claim the measure of A_ε is zero for any $\varepsilon > 0$. Suppose not, so $|A_\varepsilon| = 2\delta > 0$ for some $\varepsilon > 0$. By Lusin's theorem, there is a compact $K \subset A_\varepsilon$ such that $|K| \ge \delta$ and $v|_K$ is continuous. Then $v(K)$ is compact and is contained in $(0,\infty)$ since v is positive at each point of $[0,\bar\varphi(t))$.

Apply Lemma 3.8 with $\eta(\varphi) = \eta_n(\varphi) = \max\{0, 1 - n\,\mathrm{dist}(\varphi, v(K))\}$ for $n = 1,2,\ldots$. Then η_n has compact support in $(0,\infty)$ for n sufficiently large and converges boundedly pointwise to the characteristic function $\chi_{v(K)}$. It follows that

$$0 = \int_0^1 \chi_{v(K)}(v(\varphi))g(\varphi)\,d\varphi.$$

Now if $v(x) \in v(K)$, then $v(x) = v(y)$ for some $y \in K$, and either $g(x) = g(y) > \varepsilon$ or $x \in \Delta$. It follows that

$$\int_0^1 \chi_{v(K)}(v(\varphi))g(\varphi)\,d\varphi \ge \varepsilon|K| > 0,$$

yielding a contradiction. Hence $|A_\varepsilon| = 0$ for any $\varepsilon > 0$. A similar argument applies for $\{x \mid g(x) < -\varepsilon\}$, and we then deduce that $g(\varphi) = 0$ for almost every $\varphi \in [0,\bar\varphi(t))$. This proves the Lemma. □

Note that for $t \in \Omega$ and $\varphi \in (\bar{\varphi}(t), 1]$ we have that $v(t,\varphi) = 0$, and $\partial_t v(t,\varphi) = 0$.

Since $(t,\varphi) \mapsto v(t,\varphi)$ is right continuous in φ and locally Lipschitz in t uniformly in φ, the set $Q = \{(t,\varphi) \mid v(t,\varphi) > 0\}$ is open in $(0,\infty) \times [0,1]$. Define

$$g(t,\varphi) = \Lambda(v(t,\varphi), \theta(t)) - \overline{\partial}_t v(t,\varphi);$$

then g is measurable on $(0,\infty) \times [0,1]$ and by Lemma 3.10 we have

$$\int_0^\infty \int_0^1 \chi_Q |g| \, d\varphi \, dt = 0.$$

By Fubini's theorem, we have $\chi_Q g = 0$ a.e. Hence there exists a set S of full measure in $[0,1]$ such that if $\varphi \in S$, then $(\chi_Q g)(t,\varphi) = 0$ for almost every t, and $t \mapsto v(t,\varphi)$ is differentiable almost everywhere.

Lemma 3.11. *If $\varphi \in S$ and $v(t,\varphi) > 0$, then $v(s,\varphi) > 0$ for all $s \in [0,t]$ and*

$$v(t,\varphi) = v(0,\varphi) + \int_0^t \Lambda(v(s,\varphi), \theta(s)) \, ds. \tag{3.8}$$

Proof. For any $t_1 \in (0,t)$, if $v(s,\varphi) > 0$ for all $s \in [t_1, t]$, then since $s \mapsto v(s,\varphi)$ is differentiable and $g(s,\varphi) = 0$ a.e. in $[t_1,t]$, we have

$$v(t,\varphi) = v(t_1,\varphi) + \int_{t_1}^t \Lambda(v(s,\varphi), \theta(s)) \, ds.$$

We claim that the set $\{t_1 \in [0,t) \mid v(s,\varphi) > 0 \text{ for all } s \in [t_1,t]\}$ has the infimum $t_* = 0$. Note that the set is nonempty by the continuity of $s \mapsto v(s,\varphi)$. Suppose the infimum t_* is positive. Then $v(t_*,\varphi) = 0 < v(s,\varphi)$ for $s \in (t_*,t]$. We know that $\theta(s)$ is bounded for $s \in [0,t]$, so for some sufficiently small $h > 0$ it follows that $\Lambda(v(s,\varphi),\theta(s)) < -1/2$ for $t_* < s < t_* + h$. Then we have

$$0 < v(t_* + h, \varphi) = 0 + \int_{t_*}^{t_*+h} \Lambda(v(s,\varphi), \theta(s)) \, ds < -h/2,$$

a contradiction. Hence our claim holds: $v(s,F) > 0$ for all $s \in [0,t]$, and the formula asserted in the lemma follows. ◻

Now we can finish the proof of Proposition 3.6. Suppose $t > 0$, $\varphi \in [0,1]$ are arbitrary and $v(t,\varphi) > 0$. Since $v(t,\cdot)$ is right continuous and decreasing, there exists a sequence of numbers $\varphi_n \in S$ such that $\varphi_n > \varphi$, $\varphi_n \to \varphi$ as $n \to \infty$ and $v(t,\varphi_n) > 0$. Using Lemma 3.11 it follows that $v(s,\varphi) > 0$ for all $s \in [0,t]$ and

$$\begin{aligned} &v(t,\varphi) - v(0,\varphi) - \int_0^t \Lambda(v(s,\varphi),\theta(s)) \, ds \\ &\quad = \lim_{n\to\infty} \left(v(t,\varphi_n) - v(0,\varphi_n) - \int_0^t \Lambda(v(s,\varphi_n),\theta(s)) \, ds \right) = 0. \end{aligned}$$

This completes the proof of Proposition 3.6. ◻

The results asserted in Theorem 1.3 now follow directly from Theorems 1.1 and 1.2, with the help of Propositions 3.5 and 3.6, Corollary 3.4, and (3.5).

4. A different conserved quantity. In the theory of Ostwald ripening, one also encounters an alternative to the condition that the total particle volume is conserved in time. If mass in the diffusion field is taken into account, one finds that a quantity of the form

$$Q = a\theta(t) + \int_0^1 v(t,\varphi)\, d\varphi \tag{4.1}$$

is conserved instead, where $a > 0$ is a constant. $\theta(t)$ need no longer be positive.

In terms of the theory developed in this paper, the constraint (4.1) is simpler to deal with than the constraint of constant volume. One has the bound

$$\theta(t) \le Q/a,$$

and when comparing two solutions of (2.2), one can use the arguments of Lemma 2.7 and replace the use of Lemma 2.9 by the simpler estimate

$$|\theta_1(t) - \theta_2(t)| \le a^{-1}\|v_1(t,\cdot) - v_2(t,\cdot)\| + a^{-1}|Q_1 - Q_2|. \tag{4.2}$$

From the standard Gronwall's inequality, one easily deduces the a priori estimate asserted in the following result. The existence and uniqueness proofs are the same as in section 2.

THEOREM 4.1. *Let $v_0 \in rcd([0,1])$, $Q \in \mathbb{R}$. Then there exists a unique function $v \in C([0,\infty), rcd([0,1]))$ such that, with $\theta(t)$ determined by* (4.1), *we have*

$$v(t,\varphi) = v_0(\varphi) + \int_0^t (v(s,\varphi)^{1/3}\theta(s) - 1)\, ds$$

whenever $v(t,\varphi) > 0$.

Given $T > 0$, $C_0 > 0$, there exists a positive constant C such that, given two solutions as above which also satisfy $\max(Q_1,Q_2) \le C_0$, *then*

$$\sup_{0\le t\le T} \|v_1(t,\cdot) - v_2(t,\cdot)\| \le C\left(\|v_1(0,\cdot) - v_2(0,\cdot)\| + |Q_1 - Q_2|\right).$$

Using the correspondence $v(t,\cdot) \mapsto \nu_t = \hat{\nu}(v(t,\cdot))$ and its inverse as in section 3, from Propositions 3.5 and 3.6 one may deduce directly the following corollary of Theorem 4.1.

THEOREM 4.2. *Given $\nu_0 \in \mathcal{P}_0$, $Q \in \mathbb{R}$, there exists a unique map $t \mapsto \nu_t$ that is locally Lipschitz from $[0,\infty)$ into $\mathcal{P}_0$ such that, with $\theta(t)$ determined by the relation*

$$Q = a\theta(t) + \int_0^\infty v\, d\nu_t(v),$$

we have

$$\int_0^\infty \int_0^\infty \partial_t\zeta(t,v) + \Lambda(v,\theta(t))\partial_v\zeta(t,v)\ d\nu_t(v)\, dt = 0$$

for all smooth $\zeta : (0,\infty)\times(0,\infty) \to \mathbb{R}$ with compact support.

Given any $T > 0$, $C_0 > 0$, there exists $C > 0$ such that, if two such weak solutions $\nu^{(1)}$, $\nu^{(2)}$ are given, which satisfy $\max(Q_1,Q_2) \le C_0$, *then*

$$\sup_{0\le t\le T} d_\infty(\nu_t^{(1)}, \nu_t^{(2)}) \le C\left(d_\infty(\nu_0^{(1)}, \nu_0^{(2)}) + |Q_1 - Q_2|\right).$$

Acknowledgments. The authors gratefully acknowledge discussions with S. Müller, B. Kirchheim, F. Otto, S. Luckhaus, G. Friesecke, and R. Kohn.

REFERENCES

[1] N. D. Alikakos and G. Fusco, *The equations of Ostwald ripening for dilute systems,* J. Statist. Phys., 95 (1999), pp. 851–866.

[2] J. Carr and O. Penrose, *Asymptotic behavior in a simplified Lifshitz-Slyozov equation,* Phys. D, 124 (1998), pp. 166–176.

[3] H. Federer, *Geometric Measure Theory,* Springer-Verlag, New York, 1969.

[4] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions,* Michigan Math. J., 31 (1984), pp. 231–240.

[5] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker-Planck equation,* SIAM J. Math. Anal., 29 (1998), pp. 1–17.

[6] I. M. Lifshitz and V. V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions,* J. Phys. Chem. Solids, 19 (1961), pp. 35–50.

[7] B. Niethammer, *Derivation of the LSW theory for Ostwald ripening by homogenization methods,* Arch. Rational Mech. Anal., 147 (1999), pp. 119–178.

[8] B. Niethammer and R. L. Pego, *Non self similar behavior in the LSW theory of Ostwald ripening,* J. Statist. Phys., 95 (1999), pp. 867–902.

[9] F. Otto, *Dynamics of labyrinthine pattern formation in magnetic fluids: A mean-field theory,* Arch. Rational Mech. Anal., 141 (1998), pp. 63–103.

[10] O. Penrose, *The Becker-Döring equations at large times and their connection with the LSW theory of coarsening,* J. Statist. Phys., 89 (1997), pp. 305–320.

[11] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models,* John Wiley, New York, 1991.

[12] J. J. L. Velázquez, *The Becker-Döring equations and the Lifshitz-Slyozov theory of coarsening,* J. Statist. Phys., 92 (1998), pp. 195–236.

[13] P. W. Voorhees, *The theory of Ostwald ripening*, J. Statist. Phys., 38 (1985), pp. 231–252.

[14] P. W. Voorhees, *Ostwald ripening of two-phase mixtures,* Ann. Rev. Mater. Sci., 22 (1992), pp. 197–215.

[15] C. Wagner, *Theorie der Alterung von Niederschlägen durch Umlösen,* Z. Elektrochem., 65 (1961), pp. 581–594.

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 486–513

GLOBAL SMALL AMPLITUDE SOLUTIONS OF NONLINEAR HYPERBOLIC SYSTEMS WITH A CRITICAL EXPONENT UNDER THE NULL CONDITION*

AKIRA HOSHIGA† AND HIDEO KUBO‡

Dedicated to Professor Rentaro Agemi on the occasion of his 60th birthday.

Abstract. This paper deals with the Cauchy problems of nonlinear hyperbolic systems in two space dimensions with small data. We assume that the propagation speeds differ from each other and that nonlinearities are cubic. Then it will be shown that if the nonlinearities satisfy the *null condition*, there exists a global smooth solution. To prove this kind of claim, one usually makes use of the generalized differential operators Ω_{ij}, S, and L_i, which will be introduced in section 1. But it is difficult to adopt the operators $L_i = x_i\partial_t + t\partial_{x_i}$ to our problem, because they do not commute with the d'Alembertian whose propagation speed is not equal to one. We succeed in taking L_i away from the proof of our theorem. One can apply our method to a scalar equation; hence L_i are needless in this kind of argument.

Key words. null condition, different speeds, a unique global smooth solution

AMS subject classifications. 35A05, 35B45, 35L15, 35L55

PII. S0036141097326064

1. Introduction and statement of main result.

We consider the initial value problem for

$$\square_i u^i \equiv \partial_t^2 u^i - c_i^2 \Delta u^i = F^i(\partial u, \partial^2 u) \quad \text{in } \mathbf{R}^n \times (0,\infty), \tag{1.1}$$

$$u^i(x,0) = \varepsilon f^i(x), \quad \partial_t u^i(x,0) = \varepsilon g^i(x) \quad \text{in } \mathbf{R}^n, \tag{1.2}$$

where $i = 1,\dots,m$, $n = 2,3,c_i$ are positive constants and $\varepsilon > 0$ is a small parameter. Besides, $F^i \in C^\infty(\mathbf{R}^{(n+1)m} \times \mathbf{R}^{(n+1)^2 m})$ and f^i, $g^i \in C_0^\infty(\mathbf{R}^n)$. We also denoted $u = (u^1,\dots,u^m)$, $\partial = (\partial_t, \partial_1, \dots, \partial_n)$ with $\partial_t = \partial/\partial t$, $\partial_j = \partial/\partial x_j$ and $\partial^2 u$ stands for the second derivatives of u. As for F^i, we assume

$$F^i(\partial u, \partial^2 u) = \sum_{l=1}^{m} \sum_{\gamma,\delta=0}^{n} H_{il}^{\gamma\delta}(\partial u)\partial_\gamma\partial_\delta u^l + K_i(\partial u), \tag{1.3}$$

where $H_{il}^{\gamma\delta}$ and $K_i \in C^\infty(\mathbf{R}^{(n+1)m})$ satisfy

$$H_{il}^{\gamma\delta}(\partial u) = O(|\partial u|^{p-1}), \quad K_i(\partial u) = O(|\partial u|^p) \quad \text{near } \partial u = 0. \tag{1.4}$$

Here p is an integer with $p > 1$. In order to derive an energy estimate we further assume

$$H_{il}^{\gamma\delta}(\partial u) = H_{li}^{\gamma\delta}(\partial u) = H_{il}^{\delta\gamma}(\partial u). \tag{1.5}$$

*Received by the editors August 15, 1997; accepted for publication (in revised form) March 16, 1999; published electronically February 1, 2000.
http://www.siam.org/journals/sima/31-3/32606.html

†Faculty of Engineering, Kitami Institute of Technology, Kitami, Hokkaido 090, Japan (a-hoshig@ptolemy.comm.kitami-it.ac.jp).

‡Department of Applied Mathematics at Ohya, Faculty of Engineering, Shizuoka University, Shizuoka, Japan.

Although our interest lies in the case where the system (1.1) has different propagation speeds, we start with a review of known results for the case where $m = 1$ or the system (1.1) has same propagation speeds. Indeed, such cases have been studied extensively. Set $p_c = (n+1)/(n-1)$. If $p > p_c$, then the problem (1.1) and (1.2) has a smooth global solution for sufficiently small ε. Moreover, if $p = p_c$, then the problem (1.1) and (1.2) admits an "almost" global solution for small initial data. (See F. John and S. Klainerman [12], S. Klainerman [16], and M. Kovalyov [19], for instance). On the other hand, if $1 < p \le p_c$, then the problem (1.1) and (1.2) does not admit global solutions in general. (See R. Agemi [1], S. Alinhac [3], L. Hörmander [7], A. Hoshiga [9], and F. John [10].) Therefore, we shall call the number p_c the critical exponent in the following.

In the critical case $p = p_c$, the following interesting result is known. If the nonlinearity has a special form, a global solution of (1.1) and (1.2) exists, instead of an almost global solution. (See D. Christodoulou [4], P. Godin [6], A. Hoshiga [8], F. John [11], S. Katayama [13], and S. Klainerman [17], for instance.) We shall call the restriction on the nonlinearlities *null condition*, according to S. Klainerman [15]. We will restrict ourselves to the case where $n = 2$ and $p = p_c = 3$. Then, when $c_1 = \cdots = c_m = 1$, the null condition is stated as follows: For any $i, j, k, l = 1, \ldots, m$,

$$(1.6) \qquad \sum_{\alpha,\beta,\gamma=0}^{2} A_{ijkl}^{\alpha\beta\gamma} X_\alpha X_\beta X_\gamma = 0 \quad \text{and} \quad \sum_{\alpha,\beta,\gamma,\delta=0}^{2} D_{ijkl}^{\alpha\beta\gamma\delta} X_\alpha X_\beta X_\gamma X_\delta = 0$$

hold on the hypersurface $(X_0)^2 - c_i^2\{(X_1)^2 + (X_2)^2\} = 0$, where we have set

$$(1.7) \ A_{ijkl}^{\alpha\beta\gamma} \equiv \left. \frac{\partial^3 K_i(\partial u)}{\partial(\partial_\alpha u^j)\partial(\partial_\beta u^k)\partial(\partial_\gamma u^l)} \right|_{\partial u=0} \quad \text{and} \quad D_{ijkl}^{\alpha\beta\gamma\delta} \equiv \left. \frac{\partial^2 H_{il}^{\gamma\delta}(\partial u)}{\partial(\partial_\alpha u^j)\partial(\partial_\beta u^k)} \right|_{\partial u=0}.$$

A role of the null condition is closely connected to the following vector fields which generate a Lie algebra with respect to the usual commutator of linear operators:

$$(1.8) \qquad \partial_t, \partial_1, \partial_2, \quad S = t\partial_t + r\partial_r, \quad \Omega = x_1\partial_2 - x_2\partial_1,$$

and

$$L_i = x_i\partial_t + t\partial_i \quad (i = 1, 2),$$

where $r = |x|$. In fact, we may write

$$(1.9) \qquad \partial_i = -\omega_i\partial_t + \frac{1}{t}L_i + \frac{\omega_i}{t+r}S - \sum_{j=1}^{2} \frac{r\omega_i\omega_j}{t(t+r)}L_j \quad (i = 1, 2),$$

where $\omega_i = x_i/|x|$. (See [11].) In the leading terms of F^i, replacing ∂_i with (1.9) and using the null condition (1.6), we get

$$(1.10) \ |\Gamma^a F^i(\partial u, \partial^2 u)| \le \frac{C}{t} \sum_{|b+c+d| \le |a|+1} |\Gamma^b u||\Gamma^c \partial u||\Gamma^d \partial u| + \text{ (higher order terms)},$$

which gives us an additional decaying factor t^{-1}. This is a crucial point to treat the critical nonlinearity.

We now turn our attention to the case where $m \ge 2$ and the propagation speeds are different from each other when $n = 2$ and $p = 3$. M. Kovalyov proved the existence of the global solution of (1.1) and (1.2) in [20] under the assumption that for each $i(=1,\ldots,m)$, $A^{\alpha\beta\gamma}_{ijjj}=0$ for any $\alpha,\beta,\gamma=0,1,2$, $j=1,\ldots,m$ and $H^{\gamma\delta}_{il}(\partial u)\equiv 0$ for any $\gamma,\delta=0,1,2$, $l=1,\ldots,m$. In [2], R. Agemi and K. Yokoyama had the same result under the weaker assumption that for each $i(=1,\ldots,m)$, $A^{\alpha\beta\gamma}_{iiii}=0$ for any $\alpha,\beta,\gamma=0,1,2$ and $D^{\alpha\beta\gamma\delta}_{iiii}=0$ for any $\alpha,\beta,\gamma,\delta=0,1,2$. Here we have used the notation in (1.7). These results imply that when the propagation speeds are distinct, the global solution of (1.1) and (1.2) exists even if the nonlinearities do not satisfy (1.6). In this paper, we would like to show more generally that when the propagation speeds are distinct, (1.1) and (1.2) has a global solution under the following condition: For each $i=1,\ldots,m$,

$$\sum_{\alpha,\beta,\gamma=0}^{2} A^{\alpha\beta\gamma}_{iiii}X_\alpha X_\beta X_\gamma=0 \quad\text{and}\quad \sum_{\alpha,\beta,\gamma,\delta=0}^{2} D^{\alpha\beta\gamma\delta}_{iiii}X_\alpha X_\beta X_\gamma X_\delta=0 \tag{1.11}$$

hold on the hypersurface $(X_0)^2-c_i^2\{(X_1)^2+(X_2)^2\}=0$. Having the condition (1.11) in mind, we shall rewrite F^i in the following form:

$$F^i(\partial u,\partial^2 u)=N^i(\partial u^i,\partial^2 u^i)+R^i(\partial u,\partial^2 u)+G^i(\partial u,\partial^2 u), \tag{1.12}$$

where

$$\begin{aligned} N^i(\partial u^i,\partial^2 u^i)&=\sum_{\alpha,\beta,\gamma,\delta=0}^{2} D^{\alpha\beta\gamma\delta}_{iiii}\partial_\alpha u^i\partial_\beta u^i\partial_\gamma\partial_\delta u^i+\sum_{\alpha,\beta,\gamma=0}^{2} A^{\alpha\beta\gamma}_{iiii}\partial_\alpha u^i\partial_\beta u^i\partial_\gamma u^i,\\ R^i(\partial u,\partial^2 u)&=\sum_{j,k,l=1}^{m}\sum_{\alpha,\beta,\gamma,\delta=0}^{2} E^{\alpha\beta\gamma\delta}_{ijkl}\partial_\alpha u^j\partial_\beta u^k\partial_\gamma\partial_\delta u^l\\ &\quad+\sum_{j,k,l=1}^{m}\sum_{\alpha,\beta,\gamma=0}^{2} B^{\alpha\beta\gamma}_{ijkl}\partial_\alpha u^j\partial_\beta u^k\partial_\gamma u^l, \end{aligned}$$

and

$$G^i(\partial u,\partial^2 u)=\sum_{l=1}^{m}\sum_{\gamma,\delta=0}^{2} H_{il}(\partial u)\partial_\gamma\partial_\delta u^l+M_i(\partial u).$$

Here $E^{\alpha\beta\gamma\delta}_{ijkl}$ and $B^{\alpha\beta\gamma}_{ijkl}$ are defined by

$$\begin{aligned} E^{\alpha\beta\gamma\delta}_{ijkl}&=\begin{cases} D^{\alpha\beta\gamma\delta}_{ijkl} & (j,k,l)\neq(i,i,i),\\ 0 & (j,k,l)=(i,i,i),\end{cases}\\ B^{\alpha\beta\gamma}_{ijkl}&=\begin{cases} A^{\alpha\beta\gamma}_{ijkl} & (j,k,l)\neq(i,i,i),\\ 0 & (j,k,l)=(i,i,i).\end{cases} \end{aligned} \tag{1.13}$$

Also, we assume H_{il} and $M_i\in C^\infty(\mathbf{R}^{3m})$ satisfy

$$H_{il}(\partial u)=O(|\partial u|^3),\quad M_i(\partial u)=O(|\partial u|^4)\quad\text{near } \partial u=0.$$

By (1.11), N^i has the usual null-form for a scalar wave equation. Its concrete form will be proposed in section 3. We shall call N^i the null-form, while R^i is the resonance-form.

Now we state our main theorem.

THEOREM 1.1. *Let $n = 2$ and $c_i \neq c_j$ if $i \neq j$. Suppose that* (1.12), (1.5), *and* (1.11) *hold. Then there exists a positive constant ε_0 such that the initial value problem* (1.1) *and* (1.2) *has a unique C^∞-solution in $\mathbf{R}^2 \times [0,\infty)$ for $0 < \varepsilon \leq \varepsilon_0$.*

Remark 1. We would like to mention here the key idea of the proof of Theorem 1.1. Compared with the case where the system (1.1) has common propagation speeds, a treatment of the null-form is much more complicated when the speeds are different. The difficulty comes from the simple fact that L_j does not commute with $\Box_i$ if $c_i \neq 1$. Therefore, it seems difficult to adopt the operator L_j (or some modification of them) for the system (1.1) with different propagation speeds. Our main idea in this paper is to use the operator S effectively. More precisely, in order to obtain a variant of (1.10) without using L_j, we shall use the following relation instead of (1.9):

$$\partial_t = -c_i\partial_r + \frac{c_i t - r}{t}\partial_r + \frac{1}{t}S \tag{1.14}$$

and

$$\nabla = \frac{x}{r}\partial_r - \frac{x^\perp}{r^2}\Omega, \tag{1.15}$$

where $\nabla = (\partial_1, \partial_2)$ and $x^\perp = (x_2, -x_1)$. Since we need an additional decaying factor only in the region near the characteristic lay, we rewrite (1.14) as

$$\partial_t = -c_i\partial_r - \frac{\delta(r,t)}{\sqrt{t}}\partial_r + \frac{1}{t}S \quad \text{for } |c_i t - r| \leq \sqrt{t}, \tag{1.16}$$

where $-1 \leq \delta(r,t) \leq 1$. This is a key point in our argument. (For the details, see section 3 below). Moreover, this approach also works when either $m = 1$ or $c_1 = \cdots = c_m$ holds.

Remark 2. The other attempts to argue within the framework of $(\partial_\alpha, \Omega, S)$ were also done by S. Klainerman and T. Sideris [18] and by T. Sideris [23]. They studied the nonlinear elastic waves with the critical exponent. They used the operator S in order to extract a decaying factor from the elastic wave operator. However, their method requires that the nonlinearity has a divergence structure. Unfortunately, we can not apply their method to our case due to the lack of such a structure. Hence, following [19], [20], and [2], we make use of L^∞-weighted estimates derived by estimating the fundamental solution of the wave operator $\partial_t^2 - \Delta$, pointwisely. (See also section 4 below.)

2. Notations. In this section we collect some notations which will be used in the following discussion. Without loss of generality, we may assume

$$c_1 > c_2 > \cdots > c_m. \tag{2.1}$$

We denote the vector fields introduced in (1.8) by Γ_i as follows:

$$\Gamma = (\Gamma_0, \Gamma_1, \ldots, \Gamma_4) = (\partial, \Omega, S),$$

where

$$\partial = (\partial_0, \partial_1, \partial_2), \quad \partial_0 = \partial_t, \quad \Omega = x_1\partial_2 - x_2\partial_1, \quad \text{and} \quad S = t\partial_t + r\partial_r.$$

We can easily verify the following commutator relations:

$$[\Gamma_\sigma, \Box_i] = -2\delta_{4\sigma}\Box_i \quad \text{for} \quad \sigma = 0, \dots, 4, i = 1, \dots, m \tag{2.2}$$

and

$$\begin{aligned} &[\partial_\alpha, \partial_\beta] = 0 \quad (\alpha, \beta = 0, 1, 2), \quad [\Omega, \partial_0] = 0, \quad [\Omega, \partial_1] = -\partial_2, \quad [\Omega, \partial_2] = \partial_1, \\ &[S, \partial_\alpha] = -\partial_\alpha \quad (\alpha = 0, 1, 2), \quad [S, \Omega] = -\Omega. \end{aligned} \tag{2.3}$$

Here $[,]$ denotes the usual commutator of linear operators and $\delta_{\alpha\beta}$ is Kronecker's delta.

Next we define several norms for a vector valued function $u(x,t)$:

$$\begin{aligned} |u(t)|_k &= \sum_{|a|\le k} \sum_{i=1}^m \|\Gamma^a u^i(\cdot, t)\|_{L^\infty}, \\ [u(t)]_k &= \sum_{|a|\le k} \sum_{i=1}^m \|w_i(|\cdot|, t)\Gamma^a u^i(\cdot, t)\|_{L^\infty}, \\ \|u(t)\|_k &= \sum_{|a|\le k} \sum_{i=1}^m \|\Gamma^a u^i(\cdot, t)\|_{L^2}, \end{aligned}$$

where k is a nonnegative integer, $a = (a_0, \dots, a_4)$ is a multi-index, $\Gamma^a = \Gamma_0^{a_0} \cdots \Gamma_4^{a_4}$, and $|a| = a_0 + \cdots + a_4$. In addition, w_i is the following weight function associated with the ith component of u:

$$w_i(r,t) = (1+r)^{\frac{1}{2}-\gamma}(1+t+r)^\gamma(1+|c_it-r|)^{\frac{1}{2}} \quad \text{for } r \ge 0, t \ge 0,$$

where $1/4 < \gamma < 1/2$. Moreover, we also use

$$|u|_{k,T} = \sup_{0<t<T} |u(t)|_k, \quad [u]_{k,T} = \sup_{0<t<T} [u(t)]_k, \quad \|u\|_{k,T} = \sup_{0<t<T} \|u(t)\|_k.$$

Next we split the region $(0,\infty) \times (0,\infty)$ for each $i (i = 1, \dots, m)$ as follows:

$$\tilde{\Lambda}_i = \left\{(r,t) \in (0,\infty) \times (0,\infty) : \frac{1}{3}\left(2 + \frac{c_i}{c_{i-1}}\right) r \le c_i t \le \frac{1}{3}\left(2 + \frac{c_i}{c_{i+1}}\right) r \ \text{ and } \ r \ge 1\right\}$$

and $\tilde{\Lambda}_i^c = ((0,\infty) \times (0,\infty)) \setminus \tilde{\Lambda}_i$, where we have set $c_0 = 4c_1$ and $c_{m+1} = c_m/4$. Because of (2.1), this definition is meaningful. In particular, we have

$$\tilde{\Lambda}_i \cap \tilde{\Lambda}_l = \emptyset \quad \text{if } i \ne l. \tag{2.4}$$

Using the fact that $1+r$ is equivalent to $1+t+r$ for $(r,t) \in \tilde{\Lambda}_i$, while, so is $1+|c_it-r|$ for $(r,t) \in \tilde{\Lambda}_i^c$, we easily see that

$$w_i(r,t) \ge C(1+t+r)^{\frac{1}{2}} \quad \text{for } (r,t) \in (0,\infty) \times (0,\infty) \tag{2.5}$$

and that if $\gamma > 1/4$,

$$w_i(r,t) \ge C(1+t+r)^{\frac{3}{4}} \quad \text{for } (r,t) \in \tilde{\Lambda}_i^c. \tag{2.6}$$

We conclude this section by showing an important property of the weight function based on the following other decomposition of $(0,\infty)\times(0,\infty)$ for each $i(i=1,\dots,m)$:

$$\Lambda_i=\{(r,t)\in(0,\infty)\times(0,\infty):|c_it-r|\le\sqrt{t}\}$$

and $\Lambda_i^c=((0,\infty)\times(0,\infty))\setminus\Lambda_i$.

Proposition 2.1. *Let $1/4<\gamma<1/2$ and $i=1,2,\dots,m$. Then we have*

$$w_i(r,t)\ge C(1+t)^{\frac{3}{4}}\quad \textit{for } (r,t)\in\Lambda_i^c, \tag{2.7}$$

$$w_i(r,t)\le C(1+t)^{\frac{3}{4}}\quad \textit{for } (r,t)\in\Lambda_i. \tag{2.8}$$

Proof. First we shall show (2.7). If $(r,t)\in\tilde{\Lambda}_i^c\cap\Lambda_i^c$, we have

$$w_i(r,t)\ge C(1+t+r)^{\gamma+\frac{1}{2}}\ge C(1+t+r)^{\frac{3}{4}}$$

for $\gamma>1/4$. If $(r,t)\in\tilde{\Lambda}_i\cap\Lambda_i^c$, we have

$$w_i(r,t)\ge C(1+t+r)^{\frac{1}{2}}(1+\sqrt{t})^{\frac{1}{2}}\ge C(1+t)^{\frac{3}{4}}.$$

We thus obtain (2.7).

Next we shall show (2.8). Note that

$$\frac{c_it}{2}\le r\le 2c_it\quad\text{for } (r,t)\in\Lambda_i\quad\text{with } t\ge\frac{4}{c_i^2}. \tag{2.9}$$

Therefore, we get

$$w_i(r,t)\le C(1+t)^{\frac{1}{2}}(1+\sqrt{t})^{\frac{1}{2}}\le C(1+t)^{\frac{3}{4}}$$

for such (r,t). On the other hand, if $(r,t)\in\Lambda_i$ and $0\le t\le 4/c_i^2$, r is also bounded by some uniform constant, hence (2.8) follows. This completes the proof. □

3. An estimate for the null-form. By (1.11), one can write N^i defined in (1.12) as linear combinations of the following:

$$\begin{aligned}
N_1^i&=((\partial_0u^i)^2-c_i^2|\nabla u^i|^2)\partial_\alpha\partial_\beta u^i,\\
N_2^i&=\partial_\alpha u^i\partial_\beta((\partial_0u^i)^2-c_i^2|\nabla u^i|^2),\\
N_3^i&=\partial_\alpha u^i\partial_\beta u^i\Box_i u^i,\\
N_4^i&=\partial_\alpha u^i(\partial_\beta u^i\partial_\gamma\partial_\delta u^i-\partial_\gamma u^i\partial_\beta\partial_\delta u^i),\\
N_5^i&=\partial_\alpha u^i((\partial_0u^i)^2-c_i^2|\nabla u^i|^2)
\end{aligned}$$

for $\alpha,\beta,\gamma,\delta=0,1,2$. As we have already discussed in introduction, we shall extract an additional decaying factor from the null-form, by making use of their special form together with the identity (1.16). This is a crucial point in our argument.

Proposition 3.1. *It holds that for $i=1,\dots,m$*

$$|\Gamma^aN^i(\partial u^i,\partial^2u^i)|\le\frac{C}{\sqrt{1+t}}\Phi_a^i+\frac{C}{\sqrt{1+t}}\Theta_a^i\quad \textit{in } \Lambda_i, \tag{3.1}$$

where we have set

$$\Phi_a^i=\sum_{|b+c+d|\le|a|+1}|\partial\Gamma^bu^i||\partial\Gamma^cu^i||\partial\Gamma^du^i|,$$

$$\Theta_a^i=\sum_{\substack{|b+c+d|\le|a|+2\\|b|,|c|,|d|\le|a|+1}}|\Gamma^bu^i||\partial\Gamma^cu^i||\partial\Gamma^du^i|.$$

Proof. It is evident that (3.1) holds for $0 \le t \le \max\{1, 4/c_i^2\}$. Therefore, we shall assume $t \ge \max\{1, 4/c_i^2\}$ in the following. For simplicity, we omit the upper index i of u^i during the proof.

First, we consider the case $N^i = N_1^i$. If we set

$$Q_1(u,v) = \partial_0 u \partial_0 v - c_i^2 \nabla u \cdot \nabla v,$$

then we may write

$$\Gamma^a N_1^i = \sum_{a'+d'=a} \binom{a}{a'} \Gamma^{a'}(Q_1(u,u))\Gamma^{d'}(\partial_\alpha \partial_\beta u). \tag{3.2}$$

By the commutator relations (2.3), we obtain

$$\Gamma Q_1(u,v) = Q_1(\Gamma u, v) + Q_1(u, \Gamma v) - 2\delta_{4\sigma} Q_1(u,v) \quad \text{for } \sigma = 0, \dots, 4.$$

Therefore, we have

$$\Gamma^{a'} Q_1(u,u) = \sum_{b+c \le a'} C_{b,c}^{a'} Q_1(\Gamma^b u, \Gamma^c u). \tag{3.3}$$

By (3.2), (3.3), and $t \ge \max\{1, 4/c_i^2\}$, it suffices to show

$$|Q_1(u,v)| \le \frac{C}{\sqrt{t}} |\partial u||\partial v| + \frac{C}{t}(|\Gamma u||\partial v| + |\partial u||\Gamma v|). \tag{3.4}$$

Setting

$$\tilde{Q}_1(u,v) = \partial_0 u \partial_0 v - c_i^2 \partial_r u \partial_r v$$

and using the formula

$$\nabla = \frac{x}{r}\partial_r - \frac{x^\perp}{r^2}\Omega, \qquad x^\perp = (x_2, -x_1), \tag{3.5}$$

we get

$$Q_1(u,v) = \tilde{Q}_1(u,v) + \frac{c_i^2}{r^2}\Omega u \, \Omega v;$$

hence

$$|Q_1(u,v)| \le |\tilde{Q}_1(u,v)| + \frac{C}{r}|\partial u||\Omega v|, \tag{3.6}$$

where we used the fact that $|\Omega u|/r \le C|\partial u|$.

If we introduce operators $S_i^\pm = \partial_t \pm c_i \partial_r$, then a simple computation yields

$$2\tilde{Q}_1(u,v) = S_i^+ u S_i^- v + S_i^- u S_i^+ v.$$

Moreover, by the formula

$$S_i^+ = \partial_t + c_i \partial_r = -\frac{\delta(r,t)}{\sqrt{t}}\partial_r + \frac{1}{t}S \quad \text{with } -1 \le \delta(r,t) \le 1 \quad \text{in } \Lambda_i, \tag{3.7}$$

we obtain

$$|\tilde{Q}_1(u,v)| \le \frac{C}{\sqrt{t}}|\partial u||\partial v| + \frac{C}{t}(|Su||\partial v| + |\partial u||Sv|) \quad \text{in} \quad \Lambda_i. \tag{3.8}$$

Thus (3.6), (3.8), and (2.9) imply (3.4).

Second, from the above argument, we immediately obtain (3.1) for the case $N^i = N_2^i$ and $N^i = N_5^i$, because of the fact that

$$N_2^i = 2Q_1(\partial_\beta u, u)\partial_\alpha u \quad \text{and} \quad N_5^i = Q_1(u,u)\partial_\alpha u.$$

Third, we consider the case $N^i = N_3^i$. It follows from (2.2) that

$$\Gamma^a \Box_i u = \sum_{b\le a} C_b \Box_i \Gamma^b u \quad \text{and} \quad \Box_i \Gamma^a u = \sum_{b \le a} C_b' \Gamma^b \Box_i u, \tag{3.9}$$

where C_b and C_b' are some constants. By (3.7), (2.9), $t \ge \max\{1, 4/c_i^2\}$, and the identity

$$\Box_i u = S_i^+ S_i^- u - \frac{c_i^2}{r^2}\Omega^2 u,$$

we obtain

$$|\Box_i u| \le \frac{C}{\sqrt{t}}|\partial^2 u| + \frac{C}{t}|\Gamma \partial u|. \tag{3.10}$$

Hence, by the first identity in (3.9) and (3.10) we have (3.1) for the case $N^i = N_3^i$.

Finally, we consider the case $N^i = N_4^i$. Using a notation

$$Q_{\alpha\beta}(u,v) = \partial_\alpha u \partial_\beta v - \partial_\beta u \partial_\alpha v, \quad \alpha, \beta = 0,1,2,$$

we can write

$$N_4^i = Q_{\beta\gamma}(u, \partial_\delta u)\partial_\alpha u.$$

Note that $Q_{\beta\alpha} = -Q_{\alpha\beta}$. Moreover, it follows from (2.3) that

$$\begin{aligned}
\partial_\eta Q_{\alpha\beta}(u,v) &= Q_{\alpha\beta}(\partial_\eta u, v) + Q_{\alpha\beta}(u, \partial_\eta v), \quad \eta = 0,1,2,\\
S Q_{\alpha\beta}(u,v) &= Q_{\alpha\beta}(Su, v) + Q_{\alpha\beta}(u, Sv) - 2Q_{\alpha\beta}(u,v),\\
\Omega Q_{01}(u,v) &= Q_{01}(\Omega u, v) + Q_{01}(u, \Omega v) - Q_{02}(u,v),\\
\Omega Q_{02}(u,v) &= Q_{02}(\Omega u, v) + Q_{02}(u, \Omega v) + Q_{01}(u,v),\\
\Omega Q_{12}(u,v) &= Q_{12}(\Omega u, v) + Q_{12}(u, \Omega v).
\end{aligned}$$

Therefore we have

$$\Gamma^a Q_{\alpha\beta}(u,v) = \sum_{\gamma,\delta=0}^{2} \sum_{b+c\le a} C_{bc}^{\gamma\delta} Q_{\gamma\delta}(\Gamma^b u, \Gamma^c v). \tag{3.11}$$

On the other hand, by (2.9), (3.5), and the formula

$$\partial_t = -\frac{r}{t}\partial_r + \frac{1}{t}S,$$

we have

$$|Q_{\alpha\beta}(u,v)| \le \frac{C}{t}(|\partial u||\Gamma v| + |\Gamma u||\partial v|). \tag{3.12}$$

Combining (3.11) and (3.12), we have (3.1) for the case $N^i = N_4^i$. This completes the proof of Proposition 3.1. ∎

4. Weighted L^∞-estimates. The aim of this section is to establish weighted L^∞-estimates of a solution $u = (u^1, \dots, u^m)$ of (1.1) and (1.2) such that $u^i \in C^\infty(\mathbf{R}^2 \times [0, T))$ and satisfies

$$[\partial u]_{k,t} \le \delta_1 \quad \text{for } 0 \le t < T, \tag{4.1}$$

where k is a nonnegative integer and $\delta_1 (0 < \delta_1 < 1)$ is a real number independent of $T > 0$. A main result of this section is the following proposition.

PROPOSITION 4.1. *Suppose that* $u = (u^1, \dots, u^m)$ *is the solution of* (1.1) *and* (1.2) *and that* (1.12) *holds. Then we have for* $(|x|, t) \in \Lambda_i^c$ *with* $t < T$ *and* $|a| \le N$

$$|w_i(|x|, t)\Gamma^a \partial u^i(x, t)| \le C_N \left(\varepsilon + [\partial u]^2_{[\frac{N+2}{2}],t} \|\partial u\|_{N+4,t}\right), \tag{4.2}$$

provided (4.1) *with* $k = [(N + 2)/2]$ *holds. Moreover, if* (1.11), (1.12), *and* (4.1) *with* $k = [(N + 4)/2]$ *hold, we have for* $(x, t) \in \mathbf{R}^2 \times [0, T)$ *and* $|a| \le N$

$$|w_i(|x|, t)\Gamma^a \partial u^i(x, t)| \le C_N \left(\varepsilon + (\varepsilon + [\partial u]^2_{[\frac{N+4}{2}],t})\|\partial u\|_{N+6,t}\right). \tag{4.3}$$

Here we take δ_1 *to be sufficiently small positive number and* C_N *denotes a positive constant independent of* T *and* δ_1.

By (3.9) and (1.1), $\Gamma^a \partial^b u^i(x, t)$ satisfies

$$\Box_i \Gamma^a \partial^b u^i(x, t) = \tilde{F}^i(\partial u, \partial^2 u) \quad \text{in } \mathbf{R}^2 \times (0, T), \tag{4.4}$$

where we have set $\tilde{F}^i(\partial u, \partial^2 u) = \sum_{d \le a} C_{a,b} \partial^b \Gamma^d F^i(\partial u, \partial^2 u)$ and a, b, and d are multi-indices. Moreover, the initial values of $\Gamma^a \partial^b u^i(x, t)$ are determined by ε, f^j, and g^j $(j = 1, \dots, m)$ by using (1.1). For instance, when $a = 0$ and $\partial^b = \partial_t$, we have

$$(\partial_t u^i)(x, 0) = \varepsilon g^i(x), \quad (\partial_t^2 u^i)(x, 0) = \varepsilon c_i^2 \Delta f^i(x) + F^i(\partial u, \partial^2 u)(x, 0).$$

We can solve the second equation with respect to $(\partial_t^2 u^i)(x, 0)$ if δ_1 is sufficiently small.

Based on this, we decompose $\Gamma^a \partial^b u(x, t)$ as follows:

$$\Gamma^a \partial^b u(x, t) = u_0(x, t) + u_1(x, t) \quad \text{with} \quad u_0 = (u_0^1, \dots, u_0^m), u_1 = (u_1^1, \dots, u_1^m), \tag{4.5}$$

where u_1^i is a solution to $\Box_i u_1^i = \tilde{F}^i(\partial u, \partial^2 u)$ with the zero initial data, while u_0^i is a solution to $\Box_i u_0^i = 0$ and $u_0^i(x, 0) = (\Gamma^a \partial^b u)(x, 0)$, $\partial_t u_0^i(x, 0) = (\partial_t \Gamma^a \partial^b u)(x, 0)$. Since $f^j(x)$, $g^j(x) \in C_0^\infty(\mathbf{R}^2)$, the initial values of u_0^i also belong to $C_0^\infty(\mathbf{R}^2)$. Therefore, when $|a| + |b| \le N$, we have

$$|u_0^i(x, t)| \le M_N \varepsilon (1 + t + r)^{-\frac{1}{2}} (1 + |c_i t - r|)^{-\frac{1}{2}} \quad \text{for} \quad (x, t) \in \mathbf{R}^2 \times [0, \infty), \tag{4.6}$$

where M_N depends on L^1-norms of f^j, g^j and their finite times derivatives. (See Lemma 1 in R. T. Glassey [5] and also Lemma 4 in [19] and [21].)

Therefore, we need to estimate only u_1^i. We may assume $c_i = 1$ without loss of generality. In the following, we shall consider the solution to an inhomogeneous wave equation $(\partial_t^2 - \Delta)u = \partial^b F$ with the zero initial data. When $F \in C^\infty(\mathbf{R}^2 \times [0, T))$, we have

$$u(x, t) = \frac{1}{2\pi} \int_{|x-y| \le t} \frac{\partial^b F(y, s)}{\sqrt{t^2 - |x - y|^2}} dy. \tag{4.7}$$

Switching to polar coordinates as $x = (r\cos\theta, r\sin\theta)$ and $y = (\lambda\cos(\theta+\psi), \lambda\sin(\theta+\psi))$ as in section 2 in [19], we have

$$\begin{aligned}u(x,t) = {} & \frac{1}{2\pi}\iint_{D'} \lambda d\lambda ds \int_{-\varphi}^{\varphi} \partial^b F(\lambda\xi, s) K_1 d\psi \\ & +\frac{1}{2\pi} H(t-r) \iint_{D''} \lambda d\lambda ds \int_{-\pi}^{\pi} \partial^b F(\lambda\xi, s) K_1 d\psi,\end{aligned} \tag{4.8}$$

where H is the Heaviside function and we have set

$$\begin{aligned}\xi &= (\cos(\theta+\psi), \sin(\theta+\psi)),\\ K_1 &= K_1(\lambda, s, \psi; r, t) = \{(t-s)^2 - r^2 - \lambda^2 + 2r\lambda\cos\psi\}^{-\frac{1}{2}},\\ \varphi &= \varphi(\lambda, s; r, t) = \arccos\left[\frac{r^2+\lambda^2-(t-s)^2}{2r\lambda}\right] \quad \text{for } (\lambda, s) \in D'.\end{aligned}$$

Moreover, the domains D' and D'' are defined as follows:

$$\begin{aligned}D' &= \{(\lambda, s) \in (0,\infty)\times(0,\infty) : 0 < s < t, \lambda_- < \lambda < \lambda_+\},\\ D'' &= \{(\lambda, s) \in (0,\infty)\times(0,\infty) : 0 < s < t-r, 0 < \lambda < \lambda_-\},\end{aligned}$$

where

$$\lambda_- = |t-s-r|, \quad \lambda_+ = t-s+r. \tag{4.9}$$

The key point to get such estimates as in Proposition 4.1 is to integrate by parts with respect to λ and s. Following [19] and [2], we shall sketch this process briefly. To begin with, we split the regions of integration D' and D'' into subregions as follows:

$$\begin{aligned}& D' = \mathit{blue} \cup \mathit{white}, \quad D'' = \mathit{black} \cup \mathit{red},\\ & \mathit{blue} = \{(s,\lambda) \in D' : \lambda_- < \lambda \le \lambda_- + \delta \text{ or } \lambda_+ - \delta \le \lambda < \lambda_+\},\\ & \mathit{black} = \{(s,\lambda) \in D' : \lambda_- - \tilde{\delta} \le \lambda < \lambda_- \text{ or } 0 < \lambda \le \tilde{\delta}\},\end{aligned} \tag{4.10}$$

where we have set $\delta = \min\{r, 1/2\}$ and $\tilde{\delta} = \min\{(t-r)/2, 1/2\}$. Notice that *white* is empty if $0 < r \le 1/2$ and that *red* is empty if $0 < t-r \le 1$.

Let $\partial^b = \partial_\alpha$ $(\alpha = 0, 1, 2)$ in (4.8). Then, according to the above decompositions, we have

$$\begin{aligned}2\pi u(x,t) = {} & \iint_{blue} \lambda d\lambda ds \int_{-\varphi}^{\varphi} (\partial_\alpha F)(\lambda\xi, s) K_1 d\psi \\ & + H\left(r - \frac{1}{2}\right) \sum_{j=0}^{1} \iint_{white} \lambda d\lambda ds \int_0^1 (\partial_\alpha F)(\lambda\Xi_j, s) K_2 d\tau \\ & + H(t-r) \iint_{black} \lambda d\lambda ds \int_{-\pi}^{\pi} (\partial_\alpha F)(\lambda\xi, s) K_1 d\psi \\ & + H(t-r-1) \iint_{red} \lambda d\lambda ds \int_{-\pi}^{\pi} (\partial_\alpha F)(\lambda\xi, s) K_1 d\psi,\end{aligned} \tag{4.11}$$

where we have changed the variable as $\psi = \Psi$ in the second term and set

$$\begin{aligned}\Psi &= \Psi(\lambda, s, \tau; r, t) = \arccos[1 - (1-\cos\varphi)\tau],\\ \Xi_j &= \Xi_j(\lambda, s, \tau; r, t) = (\cos(\theta + (-1)^j\Psi), \sin(\theta + (-1)^j\Psi)),\\ K_2 &= K_2(\lambda, s, \tau; r, t) = \{2r\lambda\tau(1-\tau)(2-(1-\cos\varphi)\tau)\}^{-\frac{1}{2}}.\end{aligned}$$

Carrying out the integration by parts in the second and fourth terms, we get the following proposition.

Proposition 4.2. *Let $u(x,t)$ be the solution to $(\partial_t^2-\Delta)u=\partial_\alpha F$ with the zero initial data. If $F\in C^\infty(\mathbf{R}^2\times[0,T))$, then $|u(t,x)|$ is dominated by the following:*

$$
\begin{aligned}
I_1(F)(x,t)&=\iint_{blue}\lambda d\lambda ds\int_{-\varphi}^{\varphi}|(\partial_\alpha F)(\lambda\xi,s)|K_1d\psi,\\
I_2(F)(x,t)&=\int_{\partial(white)}\lambda d\sigma\int_0^1|F(\lambda\Xi_j,s)|K_2d\tau,\\
I_3(F)(x,t)&=\iint_{white}d\lambda ds\int_0^1\{|F(\lambda\Xi_j,s)|+|(\Omega F)(\lambda\Xi_j,s)|\}K_2d\tau,\\
I_4(F)(x,t)&=\iint_{white}\lambda d\lambda ds\int_0^1|F(\lambda\Xi_j,s)|\{|\partial_sK_2|+|\partial_\lambda K_2|\}d\tau,\\
I_5(F)(x,t)&=\iint_{white}\lambda d\lambda ds\int_0^1|(\Omega F)(\lambda\Xi_j,s)|K_2\{|\partial_s\Psi|+|\partial_\lambda\Psi|\}d\tau,\\
J_1(F)(x,t)&=\iint_{black}\lambda d\lambda ds\int_{-\pi}^{\pi}|(\partial_\alpha F)(\lambda\xi,s)|K_1d\psi,\\
J_2(F)(x,t)&=\int_{\partial(red)}\lambda d\sigma\int_{-\pi}^{\pi}|F(\lambda\xi,s)|K_1d\psi,\\
J_3(F)(x,t)&=\iint_{red}d\lambda ds\int_{-\pi}^{\pi}\{|F(\lambda\xi,s)|+|(\Omega F)(\lambda\xi,s)|\}K_1d\psi,\\
J_4(F)(x,t)&=\iint_{red}\lambda d\lambda ds\int_{-\pi}^{\pi}|F(\lambda\xi,s)|\{|\partial_sK_1|+|\partial_\lambda K_1|\}d\psi.
\end{aligned}
$$

Proof. It is easy to see that the first and second terms in (4.11) are dominated by $I_1(F)$ and $J_1(F)$, respectively. Since

$$
(\nabla F)(\lambda\xi,s)=\xi\partial_\lambda(F(\lambda\xi,s))-\frac{\xi^\perp}{\lambda}(\Omega F)(\lambda\xi,s),\quad \xi^\perp=\sin(\theta+\psi),-\cos(\theta+\psi)),
$$

we find that the fourth term in (4.11) is dominated by $J_j(F)$ $(j=2,3,4)$ by integration by parts.

To deal with the second term in (4.11), we use the following identities:

$$
\begin{aligned}
&(\partial_sF)(\lambda\Xi_j,s)=\partial_s(F(\lambda\Xi_j,s))-(-1)^j\partial_s\Psi(\Omega F)(\lambda\Xi_j,s),\\
&\quad(\nabla F)(\lambda\Xi_j,s)=\Xi_j(\partial_\lambda(F(\lambda\Xi_j,s))-(-1)^j\partial_\lambda\Psi(\Omega F)(\lambda\Xi_j,s))-\frac{\Xi_j^\perp}{\lambda}(\Omega F)(\lambda\Xi_j,s),
\end{aligned}
$$

where $\Xi_j^\perp=(\sin(\theta+(-1)^j\Psi),-\cos(\theta+(-1)^j\Psi))$. Again by integration by parts, we find that the second term is dominated by $I_j(F)$ $(j=2,\dots,5)$. The proof is complete. □

We shall use the following estimates of K_1 and K_2. For the proof, see Proposition 2.1 in [19] and also Proposition 5.3 in [2].

Lemma 4.1. *It holds that for $(\lambda,s)\in D'$*

$$
(4.12)\quad \int_{-\varphi}^{\varphi}K_1d\psi=2\int_0^1K_2d\tau\le\frac{C}{(r\lambda)^{\frac12}}\log\left[2+\frac{r\lambda}{(\lambda-\lambda_-)(\lambda_++\lambda)}H(t-s-r)\right],
$$

$$\int_0^1 \{|\partial_s K_2| + |\partial_\lambda K_2|\} d\tau \le \frac{C}{(r\lambda)^{\frac{1}{2}}(\lambda + s + r - t)}, \tag{4.13}$$

$$\int_0^1 K_2\{|\partial_s \Psi| + |\partial_\lambda \Psi|\} d\tau \le \frac{C(r+\lambda)}{\{r\lambda(\lambda^2 - \lambda_-^2)(\lambda_+^2 - \lambda^2)\}^{\frac{1}{2}}} \tag{4.14}$$

and that for $(\lambda, s) \in D''$

$$\int_{-\pi}^{\pi} K_1 d\psi \le C\{(\lambda + \lambda_-)(\lambda_+ - \lambda)\}^{-\frac{1}{2}} \log\left[2 + \frac{r\lambda}{(\lambda_- - \lambda)(\lambda_+ + \lambda)}\right], \tag{4.15}$$

$$\int_{-\pi}^{\pi} \{|\partial_s K_1| + |\partial_\lambda K_1|\} d\psi \le \frac{C}{(\lambda_- - \lambda)\{(\lambda + \lambda_-)(\lambda_+ - \lambda)\}^{\frac{1}{2}}}. \tag{4.16}$$

Now we are in a position to derive a new weighted $L^\infty - L^\infty$ estimate for the solution ∂u of (1.1) and (1.2). We introduce the following weight functions:

$$\frac{1}{\overline{w}_i(r,t)} = \frac{1}{(1+r)^{1-2\gamma}(1+t+r)^{1+2\gamma}} + \sum_{j \ne i} \frac{1}{(1+t+r)(1+|c_j t - r|)} + \frac{1}{(1+t+r)^{1+\mu}(1+|c_i t - r|)^{1-\mu}} \tag{4.17}$$

and

$$\frac{1}{\tilde{w}(r,t)} = \frac{1}{(1+r)^{1-2\gamma}(1+t+r)^{1+2\gamma}} + \sum_{j=1}^{m} \frac{1}{(1+t+r)(1+|c_j t - r|)}, \tag{4.18}$$

where $1/4 < \gamma < 1/2$ and $0 < \mu < 1$.

Proposition 4.3. *Let* u_1^i *be the solution to* $(\partial_t^2 - \Delta)u_1^i = \partial_\alpha F^i(\partial u, \partial^2 u)$ *with the zero initial data. Here* u *is a solution of* (1.1) *and* (1.2).

(i) *Let* $(r,t) \in \tilde{\Lambda}_i^c$ *with* $r = |x|$ *and* $t < T$. *Assume that* $w(r,t)$ *satisfies*

$$0 < \frac{1}{w(r,t)} \le \frac{C}{\tilde{w}(r,t)}. \tag{4.19}$$

Then we have

$$w_i(r,t)|u_1^i(x,t)| \le CM_{0,1}, \tag{4.20}$$

where we have set for a nonnegative integer k

$$M_{0,k} = \sum_{|a| \le k} \sup_{0<s<t} \sup_{y \in \mathbf{R}^2} \||y|^{\frac{1}{2}} w(|y|, s)\Gamma^a F^i(y,s)|.$$

(ii) *Let* $(x,t) \in \mathbf{R}^2 \times [0,T)$. *Assume* $\eta_j(r,t)$ $(j = 1,2)$ *satisfy*

$$0 < \frac{1}{\eta_1(r,t)} \le \frac{C}{\overline{w}_i(r,t)}, \quad 0 < \frac{1}{\eta_2(r,t)} \le \frac{C}{\tilde{w}(r,t)}. \tag{4.21}$$

Then we have

$$w_i(r,t)|u_1^i(x,t)| \le C(M_{1,1} + M_{2,1} + M_{3,1}), \tag{4.22}$$

where we have set for a positive integer k

$$\begin{aligned}
M_{1,k} &= \sum_{|a|\le k} \sup_{0<s<t} \sup_{y\in\mathbf{R}^2} \| y|^{\frac12}\eta_1(|y|,s)\Gamma^a(R^i(y,s)+G^i(y,s))|,\\
M_{2,k} &= \sum_{|a|\le k} \sup_{0<s<t} \sup_{y\in\mathbf{R}^2} \| y|^{\frac12}\eta_2(|y|,s)\Gamma^a N^i(y,s)|,\\
M_{3,k} &= \sum_{|a|\le k} \sup_{(|y|,s)\in\Lambda_i, s<t} \| y|^{\frac12}\eta_2(|y|,s)(1+s)^{\frac12}\Gamma^a N^i(y,s)|.
\end{aligned}$$

Here, we have divided the function F^i into three parts: G^i, R^i, and N^i as in (1.12).

Proof. Employing Proposition 4.2, we find that $|u_1^i(t,x)|$ is dominated by $I_j(F^i)$ $(j=1,\dots,5)$ and $J_j(F^i)$ $(j=1,\dots,4)$. □

In the proof of Proposition 5.4 in [2], the following estimates are shown. Strictly speaking, they proved only the former part of Lemma 4.2 below. However, following their proof, we find that the assumption (4.19) is sufficient to derive (4.24) with $j=3,4$ and (4.25).

Lemma 4.2. *Set*

$$\begin{aligned}
I_1' &= \iint_{blue} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\lambda ds \int_{-\varphi}^{\varphi} K_1 d\psi,\\
I_2' &= \int_{\partial(white)} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\sigma \int_0^1 K_2 d\tau,\\
I_3' &= \iint_{white} \frac{1}{\lambda^{\frac12}w(\lambda,s)} d\lambda ds \int_0^1 K_2 d\tau,\\
I_4' &= \iint_{white} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\lambda ds \int_0^1 \{|\partial_s K_2|+|\partial_\lambda K_2|\} d\tau,\\
I_5' &= \iint_{white} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\lambda ds \int_0^1 K_2\{|\partial_s \Psi|+|\partial_\lambda \Psi|\} d\tau,\\
I_1'' &= \iint_{black} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\lambda ds \int_{-\pi}^{\pi} K_1 d\psi,\\
I_2'' &= \int_{\partial(red)} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\sigma \int_{-\pi}^{\pi} K_1 d\psi,\\
I_3'' &= \iint_{red} \frac{1}{\lambda^{\frac12}w(\lambda,s)} d\lambda ds \int_{-\pi}^{\pi} K_1 d\psi,\\
I_4'' &= \iint_{red} \frac{\lambda^{\frac12}}{w(\lambda,s)} d\lambda ds \int_{-\pi}^{\pi} \{|\partial_s K_1|+|\partial_\lambda K_1|\} d\psi.
\end{aligned}$$

Assume $w(r,t)$ satisfies

$$0<\frac{1}{w(r,t)} \le \frac{C}{\overline{w}_i(r,t)}. \tag{4.23}$$

Then we have for $(x,t)\in\mathbf{R}^2\times(0,\infty)$

$$w_i(r,t)I_j' \le C, \tag{4.24}$$

$$w_i(r,t)I_j'' \le C. \tag{4.25}$$

Moreover, (4.24) *with* $j = 3, 4$ *and* (4.25) *are still true, if* $w(r,t)$ *satisfies* (4.19).

First, we shall show the statement (i) in Proposition 4.3. By the definition of $M_{0,1}$, (4.19), and Lemma 4.2, we get for $(x,t) \in \mathbf{R}^2 \times (0,\infty)$

$$w_i(r,t)I_j(F^i)(x,t) \le CM_{0,1} \quad \text{for } j = 3, 4,$$
$$w_i(r,t)J_j(F^i)(x,t) \le CM_{0,1} \quad \text{for } j = 1, \dots, 4.$$

Therefore, our task becomes to prove

$$w_i(r,t)I_j(F^i)(x,t) \le CM_{0,1} \quad \text{for } j = 1, 2, 5, \tag{4.26}$$

provided (4.19) and $(r,t) \in \tilde{\Lambda}_i^c$. Since the treatment of $I_2(F^i)$ is similar to that of $I_1(F^i)$, we shall deal with only $I_1(F^i)$ and $I_5(F^i)$. If we set

$$\frac{1}{\xi(\lambda,s)} = \frac{1}{(1+s+\lambda)(1+|s-\lambda|)}, \tag{4.27}$$

then we have from (4.17) and (4.18)

$$\frac{1}{\tilde{w}(\lambda,s)} \le \frac{1}{\overline{w}_i(\lambda,s)} + \frac{1}{\xi(\lambda,s)}. \tag{4.28}$$

Hence, using (4.24) with $j = 1, 5$, (4.12), and (4.14), we have

$$I_j(F^i)(x,t) \le CM_{0,1}(\{w_i(r,t)\}^{-1} + \tilde{I}_j(\xi)) \quad \text{for } j = 1, 5,$$

where we have set

$$\tilde{I}_1(w) = \frac{1}{r^{\frac{1}{2}}} \iint_{blue} \frac{1}{w(\lambda,s)} \log\left[2 + \frac{r\lambda}{(\lambda-\lambda_-)(\lambda_+ + \lambda)} H(t-s-r)\right] d\lambda ds, \tag{4.29}$$

$$\tilde{I}_5(w) = \frac{1}{r^{\frac{1}{2}}} \iint_{white} \frac{1}{w(\lambda,s)} \frac{r+\lambda}{\{(\lambda^2-\lambda_-^2)(\lambda_+^2-\lambda^2)\}^{\frac{1}{2}}} d\lambda ds. \tag{4.30}$$

In the following, we shall prove for $(r,t) \in \tilde{\Lambda}_i^c$

$$\tilde{I}_j(\xi) \le \frac{C}{(1+r)^{\frac{1}{2}}(1+t+r)^{\frac{1}{2}+\gamma}} \quad \text{for } j = 1, 5. \tag{4.31}$$

First we consider $\tilde{I}_1(\xi)$. It follows from (5.33) and (5.34) in [2] that for $0 \le s \le t$

$$\int_{\lambda_-}^{\lambda_-+\delta} \log\left[2 + \frac{r\lambda}{(\lambda-\lambda_-)(\lambda_+ + \lambda)} H(t-s-r)\right] d\lambda \le C\delta, \tag{4.32}$$

$$\int_{\lambda_+-\delta}^{\lambda_+} \log\left[2 + \frac{r\lambda}{(\lambda-\lambda_-)(\lambda_+ + \lambda)} H(t-s-r)\right] d\lambda \le C\delta^{\frac{1}{2}} \log[2+|t-r|]. \tag{4.33}$$

Therefore we have

$$\begin{aligned}\tilde{I}_1(\xi) &\le \frac{C\delta^{\frac{1}{2}}}{r^{\frac{1}{2}}} \log[2+|t-r|] \left\{\int_0^t \frac{1}{\xi(\lambda_-,s)} ds + \int_0^t \frac{1}{\xi(\lambda_+,s)} ds\right\} \\ &\le \frac{C\log[2+|t-r|]}{(1+r)^{\frac{1}{2}}(1+t+r)} \left\{\int_0^t \frac{1}{1+|2s-t+r|} ds + \int_0^t \frac{1}{1+|2s-t-r|} ds\right\}\end{aligned}$$

because $\delta r^{-1} \le C(1+r)^{-1}$ and $1+|t-r|$ is equivalent to $1+t+r$ for $(r,t) \in \tilde{\Lambda}_i^c$. Since $\gamma < 1/2$, we thus obtain (4.31) for $j=1$.

Next we consider $\tilde{I}_5(\xi)$. Notice that $\delta = 1/2$ if the domain *white* is not empty, hence r is equivalent to $1+r$. Moreover, since

$$\lambda \pm \lambda_- \ge \delta, \quad \lambda \pm \lambda_- \ge \delta \quad \text{for } (\lambda, s) \in \ \textit{white},$$

we have

$$\begin{aligned}\tilde{I}_5(\xi) \le \frac{C}{(1+r)^{\frac{1}{2}}} \iint_{D'} & \frac{1}{\xi(\lambda,s)} \\ & \times \frac{r+\lambda}{\{(\lambda-\lambda_-+1)(\lambda+\lambda_-+1)(\lambda_+-\lambda+1)(\lambda_++\lambda+1)\}^{\frac{1}{2}}} d\lambda ds.\end{aligned}$$

Note that

$$\begin{aligned}&\frac{r+\lambda}{\{(\lambda-\lambda_-+1)(\lambda+\lambda_-+1)(\lambda_+-\lambda+1)(\lambda_++\lambda+1)\}^{\frac{1}{2}}} \\ &\le \frac{2}{(\lambda-\lambda_-+1)^{\frac{1}{2}}} \left\{ \frac{1}{(\lambda+\lambda_-+1)^{\frac{1}{2}}} + \frac{1}{(\lambda_+-\lambda+1)^{\frac{1}{2}}} \right\},\end{aligned}$$

which follows from

$$\begin{aligned}&\lambda_+ + \lambda \ge \max\{r, \lambda\}, \\ &\lambda_+ - \lambda \ge \max\{r, \lambda\} \quad \text{for } \lambda \le \frac{\lambda_+ - \lambda_-}{2}, \\ &\lambda + \lambda_- \ge \max\{r, \lambda\} \quad \text{for } \lambda \ge \frac{\lambda_+ - \lambda_-}{2}.\end{aligned}$$

Therefore we have

$$\begin{aligned}(1+r)^{\frac{1}{2}} \tilde{I}_5(\xi) \le C \iint_{D'} & \frac{1}{\xi(\lambda,s)} \frac{1}{\{(\lambda-t+s+r+1)(\lambda+t-s-r+1)\}^{\frac{1}{2}}} d\lambda ds \\ +C \iint_{D'} & \frac{1}{\xi(\lambda,s)} \frac{1}{\{(\lambda-t+s+r+1)(t-s+r-\lambda+1)\}^{\frac{1}{2}}} d\lambda ds \\ +C \iint_{D'} & \frac{1}{\xi(\lambda,s)} \frac{1}{\{(\lambda+t-s-r+1)(t-s+r-\lambda+1)\}^{\frac{1}{2}}} d\lambda ds.\end{aligned} \tag{4.34}$$

We shall show in the following that the right-hand side of (4.34) is dominated by $C(1+t+r)^{-\gamma-1/2}$. Since

$$1+s+\lambda \ge 1+|t-r| \quad \text{for } (\lambda, s) \in D', \tag{4.35}$$

the second term is dominated by

$$\begin{aligned}&\frac{C}{1+|t-r|} \iint_{D'} \frac{1}{1+|s-\lambda|} \left\{ \frac{1}{\lambda+s-t+r+1} + \frac{1}{t+r-s-\lambda+1} \right\} d\lambda ds \\ &\le \frac{C}{2(1+t+r)} \int_{|t-r|}^{t+r} \left\{ \frac{1}{\alpha-t+r+1} + \frac{1}{t+r-\alpha+1} \right\} d\alpha \int_{-\alpha}^{t-r} \frac{1}{1+|\beta|} d\beta,\end{aligned}$$

where we have changed the variables as

$$\alpha = s+\lambda, \quad \beta = s-\lambda. \tag{4.36}$$

Since the double integral is dominated by $C\{\log(1+t+r)\}^2$, we get the desired estimate.

To treat the first and third terms, we divide the domain D' into two parts:

$$D_- = \left\{(\lambda,s) \in D' : |\lambda - s| \le \frac{1}{2}|t-r|\right\}, \quad D_-^c = white \setminus D_-. \tag{4.37}$$

Since D_- is empty if $0 < t \le r$, we have

$$\lambda + t - s - r \ge \frac{1}{2}|t-r| \quad \text{for } (\lambda,s) \in D_-. \tag{4.38}$$

On the other hand, we have

$$1 + |s-\lambda| \ge \frac{1}{2}(1+|t-r|) \quad \text{for } (\lambda,s) \in D_-^c. \tag{4.39}$$

Using these estimates together with (4.35) and changing the variables as (4.36), we find that the first term is majored by

$$\begin{aligned}
&\frac{C}{1+|t-r|}\left\{\iint_{D'} \frac{1}{(1+s+\lambda)^{\frac{1}{2}}(1+|s-\lambda|)(\lambda+s-t+r+1)^{\frac{1}{2}}}d\lambda ds\right. \\
&\quad \left. + \iint_{D'} \frac{1}{(1+s+\lambda)}\left\{\frac{1}{\lambda-t+s+r+1} + \frac{1}{\lambda+t-s-r+1}\right\}d\lambda ds\right\} \\
&\le \frac{C}{1+t+r}\left\{\int_{|t-r|}^{t+r}\left\{\frac{1}{1+\alpha} + \frac{1}{\alpha-t+r+1}\right\}d\alpha \int_{-\alpha}^{t-r}\frac{1}{1+|\beta|}d\beta\right. \\
&\quad + \int_{|t-r|}^{t+r}\frac{1}{(1+\alpha)}\frac{1}{\alpha-t+r+1}d\alpha\int_{-\alpha}^{t-r}d\beta \\
&\quad \left. + \int_{|t-r|}^{t+r}\frac{1}{(1+\alpha)}d\alpha\int_{-\alpha}^{t-r}\frac{1}{-\beta+t-r+1}d\beta\right\},
\end{aligned}$$

which yields the desired estimate. Since the third term is dealt with similarly, we omit the details. This completes the proof of (4.26).

Second, we shall show the statement (ii). By (4.21), we have for $|a| \le 1$

$$|\Gamma^a F^i(y,s)| \le \frac{M_{1,1}+M_{2,1}}{\lambda^{\frac{1}{2}}\tilde{w}(\lambda,s)}.$$

Therefore, by Lemma 4.2, we get for $(x,t) \in \mathbf{R}^2 \times (0,\infty)$

$$\begin{aligned}
w_i(r,t)I_j(F^i)(x,t) &\le C(M_{1,1}+M_{2,1}) \quad \text{for } j = 3,4, \\
w_i(r,t)J_j(F^i)(x,t) &\le C(M_{1,1}+M_{2,1}) \quad \text{for } j = 1,\dots,4.
\end{aligned}$$

Moreover, similarly to (4.26), we get for $(|x|,t) \in \tilde{\Lambda}_i^c$

$$w_i(r,t)I_j(F^i)(x,t) \le C(M_{1,1}+M_{2,1}) \quad \text{for } j = 1,2,5.$$

Thus it suffices to prove

$$w_i(r,t)I_j(F^i)(x,t) \le C(M_{1,1}+M_{2,1}+M_{3,1}) \quad \text{for } j = 1,2,5, \tag{4.40}$$

provided (4.21) and $(r,t)\in\tilde{\Lambda}_i$.

Having (3.1) in mind, we introduce a characteristic function of Λ_i denoted by $\chi(\lambda,s)$. Then we may write

$$\Gamma^a F^i=\Gamma^a(R^i+G^i)+(1-\chi)\Gamma^a N^i+\chi\Gamma^a N^i$$

and find from (4.28) and the definition of $M_{i,1}$ given in (4.22) that

$$|\Gamma^a F^i(y,s)|\le C(M_{1,1}+M_{2,1}+M_{3,1})\lambda^{-\frac{1}{2}}\left(\frac{1}{\bar{w}_i(\lambda,s)}+\frac{1-\chi(\lambda,s)}{\xi(\lambda,s)}+\frac{\chi(\lambda,s)}{\xi(\lambda,s)(1+s)^{\frac{1}{2}}}\right)$$

for $|a|\le 1$. Therefore, using (4.24), we have for $j=1,5$

$$I_j(F^i)(r,t)\le C(M_{1,1}+M_{2,1}+M_{3,1})(\{w_i(r,t)\}^{-1}+\tilde{I}_j(\tilde{\xi})), \tag{4.41}$$

where $\tilde{I}_j$ is defined in (4.29), (4.30) and we have set

$$\frac{1}{\tilde{\xi}(\lambda,s)}=\frac{1-\chi(\lambda,s)}{\xi(\lambda,s)}+\frac{\chi(\lambda,s)}{\xi(\lambda,s)(1+s)^{\frac{1}{2}}}.$$

In the following, we shall show for $(r,t)\in\tilde{\Lambda}_i$ and $j=1,5$

$$\tilde{I}_j(\tilde{\xi})\le\frac{C}{(1+r)^{\frac{1}{2}}(1+|t-r|)^{\frac{1}{2}}} \tag{4.42}$$

because $1+r$ is equivalent to $1+t+r$ for $(r,t)\in\tilde{\Lambda}_i$.

First we consider $\tilde{I}_1(\tilde{\xi})$. Using (4.32) and (4.33), we have

$$\tilde{I}_1(\tilde{\xi})\le\frac{C\log[2+|t-r|]}{(1+r)^{\frac{1}{2}}}\left\{\int_0^t\frac{1}{\tilde{\xi}(\lambda_-,s)}ds+\int_0^t\frac{1}{\tilde{\xi}(\lambda_+,s)}ds\right\},$$

because $\delta r^{-1}\le C(1+r)^{-1}$. Since

$$(1+|s-\lambda|)^{\frac{1}{4}}\ge(1+s)^{\frac{1}{8}}\quad\text{for }(\lambda,s)\in\operatorname{supp}\{1-\chi\}, \tag{4.43}$$

we have from (4.27)

$$\frac{1}{\tilde{\xi}(\lambda,s)}\le\frac{2}{(1+s+\lambda)^{\frac{3}{4}}(1+|s-\lambda|)^{\frac{3}{4}}(1+s)^{\frac{3}{8}}}.$$

Therefore, we get

$$\begin{aligned}(1+r)^{\frac{1}{2}}\tilde{I}_1(\tilde{\xi})\le{}&\frac{C\log[2+|t-r|]}{(1+|t-r|)^{\frac{3}{4}}}\\&\times\left\{\int_0^t\frac{1}{(1+|2s-t+r|)^{\frac{3}{4}}(1+s)^{\frac{3}{8}}}ds+\int_0^t\frac{1}{(1+|2s-t-r|)^{\frac{3}{4}}(1+s)^{\frac{3}{8}}}ds\right\},\end{aligned}$$

which yields (4.42) for $j=1$.

Next we consider $\tilde{I}_5(\tilde{\xi})$. From (4.34), we have

$$\begin{aligned}(1+r)^{\frac{1}{2}}\tilde{I}_5(\tilde{\xi})\le{}&C\iint_{D'}\frac{1}{\tilde{\xi}(\lambda,s)}\frac{1}{\{(\lambda-t+s+r+1)(\lambda+t-s-r+1)\}^{\frac{1}{2}}}d\lambda ds\\&+C\iint_{D'}\frac{1}{\tilde{\xi}(\lambda,s)}\frac{1}{\{(\lambda-t+s+r+1)(t-s+r-\lambda+1)\}^{\frac{1}{2}}}d\lambda ds\\&+C\iint_{D'}\frac{1}{\tilde{\xi}(\lambda,s)}\frac{1}{\{(\lambda+t-s-r+1)(t-s+r-\lambda+1)\}^{\frac{1}{2}}}d\lambda ds.\end{aligned} \tag{4.44}$$

We shall show that the right-hand side is dominated by $C(1+|t-r|)^{-1/2}$. Using (4.27) and (4.35), we have

$$\begin{aligned}\frac{1}{\tilde{\xi}(\lambda,s)} &\le \frac{2}{\xi(\lambda,s)}\\ &\le \frac{2}{(1+|t-r|)^{\frac12}(1+s+\lambda)^{\frac14}(1+|s-\lambda|)^{\frac54}}.\end{aligned}$$

Therefore, the second term is majored by $C(1+|t-r|)^{-1/2}$ times

$$\begin{aligned}&\iint_{D'} \frac{1}{(1+s+\lambda)^{\frac14}(1+|s-\lambda|)^{\frac54}}\left\{\frac{1}{\lambda+s-t+r+1}+\frac{1}{t+r-s-\lambda+1}\right\}d\lambda ds\\ &\le C\int_{|t-r|}^{t+r}\left\{\frac{1}{(1+\alpha)^{\frac54}}+\frac{1}{(\alpha-t+r+1)^{\frac54}}+\frac{1}{(t+r-\alpha+1)^{\frac54}}\right\}d\alpha\int_{-\alpha}^{t-r}\frac{1}{(1+|\beta|)^{\frac54}}d\beta,\end{aligned}$$

which yields the desired estimate.

Next we deal with the first term, by dividing the domain D' as in (4.37). Using (4.38) and (4.39), we find that the first term is majored by $C(1+|t-r|)^{-1/2}$ times

$$\begin{aligned}&\iint_{D'}\frac{1}{\xi(\lambda,s)}\frac{1}{(\lambda+s-t+r+1)^{\frac12}}d\lambda ds\\ &+\iint_{D'}\frac{(1+|s-\lambda|)^{\frac12}}{\xi(\lambda,s)}\frac{1}{\lambda+s-t+r+1}d\lambda ds\\ &+\iint_{D'}\frac{(1+|s-\lambda|)^{\frac12}}{\tilde{\xi}(\lambda,s)}\frac{1}{\lambda-s+t-r+1}d\lambda ds.\end{aligned}$$

Analogous to the above calculation, we see that the first and second terms are bounded by some constant. Since we have by (4.43)

$$\frac{(1+|s-\lambda|)^{\frac12}}{\tilde{\xi}(\lambda,s)} \le \frac{2}{(1+s)^{\frac98}(1+|s-\lambda|)^{\frac14}},$$

the third term is dominated by

$$\int_0^\infty \frac{ds}{(1+s)^{\frac98}}\int_{-\infty}^{\infty}\left\{\frac{1}{(1+|s-\lambda|)^{\frac54}}+\frac{1}{(\lambda-s+t-r+1)^{\frac54}}\right\}d\lambda \le C;$$

hence we obtain the desired estimate of the first term in the right-hand side of (4.44). Since the third term in the right-hand side of (4.44) is dealt with similarly, we omit the details. This completes the proof of the proposition.

In our analysis, we need an upper bound of not only ∂u^i but also u^i itself.

PROPOSITION 4.4. *Let u_1^i be the solution to $(\partial_t^2-\Delta)u_1^i = F^i(\partial u, \partial^2 u)$ with the zero initial data. Here u is a solution to* (1.1) *and* (1.2). *Let $0 \le \mu < 1/2$. Assume that $w(r,t)$ satisfies* (4.19). *Then we have for $(|x|,t) \in \tilde{\Lambda}_i$ with $t<T$*

$$(1+t+r)^{\mu}|u_1^i(x,t)| \le CM_{0,0}, \tag{4.45}$$

where $M_{0,0}$ is defined in (4.20).

Proof. It follows from (4.8) with $b = 0$, (4.12), (4.15), and (4.19) that

$$|u_1^i(x,t)| \le CM_{0,0}(P_1 + P_2), \tag{4.46}$$

where we have set

$$P_1 = \frac{1}{r^{\frac{1}{2}}} \iint_{D'} \frac{1}{\tilde{w}(\lambda, s)} \log \left[2 + \frac{r\lambda}{(\lambda - \lambda_-)(\lambda_+ + \lambda)} H(t - s - r) \right] d\lambda ds$$

and

$$P_2 = H(t - r) \iint_{D''} \frac{\lambda^{\frac{1}{2}}}{\tilde{w}(\lambda, s)\{(\lambda + \lambda_-)(\lambda_+ - \lambda)\}^{\frac{1}{2}}} \log \left[2 + \frac{r\lambda}{(\lambda_- - \lambda)(\lambda_+ + \lambda)} \right] d\lambda ds.$$

Since t is equivalent to r for $(r,t) \in \tilde{\Lambda}_i$, it suffices to show

$$P_j \le C(1 + r)^{-\mu} \quad \text{for } j = 1, 2. \tag{4.47}$$

First, we treat P_1. We split the domain D' into *blue* and *white* defined by (4.10). According to this decomposition, we shall write $P_1 = P_{1,blue} + P_{1,white}$. By (4.32) and (4.33), we have

$$\begin{aligned} P_{1,blue} &\le \frac{C\delta^{\frac{1}{2}}}{r^{\frac{1}{2}}} \log[2 + |t - r|] \left\{ \int_0^t \frac{1}{\tilde{w}(\lambda_-, s)} ds + \int_0^t \frac{1}{\tilde{w}(\lambda_+, s)} ds \right\} \\ &\le \frac{C \log[2 + |t - r|]}{(1 + r)^{\frac{1}{2}}} \int_0^t \frac{1}{1 + s} ds \\ &\le \frac{C}{(1 + r)^{\mu}} \end{aligned}$$

for $0 \le \mu < 1/2$ because $1 + r$ is equivalent to $1 + t + r$ for $(r,t) \in \tilde{\Lambda}_i$.

On the other hand, we have for $(\lambda, s) \in$ *white* with $0 \le s \le t - r$

$$\frac{r\lambda}{(\lambda - \lambda_-)(\lambda_+ + \lambda)} \le \frac{\lambda}{\lambda - \lambda_-} \le 1 + 2\lambda_- \le 1 + 2(t - r)$$

because $\delta = 1/2$, if *white* is not empty. Therefore we have

$$\begin{aligned} (1 + r)^{\frac{1}{2}} P_{1,white} &\le C \log[2 + |t - r|] \iint_{D'} \frac{1}{\tilde{w}(\lambda, s)} d\lambda ds \\ &\le C \log[2 + |t - r|] \int_0^t \frac{ds}{1 + s} \int_0^{t+r} \left\{ \frac{1}{1 + \lambda} + \sum_{j=1}^m \frac{1}{1 + |c_j s - \lambda|} \right\} d\lambda; \end{aligned}$$

hence $P_{1,blue} \le C(1 + r)^{-\mu}$ for $0 \le \mu < 1/2$. We thus get (4.47) for $j = 1$.

Second, we deal with P_2. Notice that $\tilde{w}(\lambda, s)$ is equivalent to $\tilde{w}(\lambda_-, s)$ for $\lambda_- - 1 \le \lambda \le \lambda_-$ and that $(\lambda_+ - \lambda)^{1/2} \le (1 + r)^{1/2}$ for $0 < \lambda \le \lambda_- - 1$ and $0 \le s \le t - r$. Moreover, we have for $0 < \lambda \le \lambda_- - 1$ and $0 \le s \le t - r$

$$\frac{r\lambda}{(\lambda_- - \lambda)(\lambda_+ + \lambda)} \le \frac{\lambda}{\lambda_- - \lambda} \le -1 + \lambda_- \le t - r.$$

Splitting the integral into two parts, we have

$$\begin{aligned}P_2 \le\ & CH(t-r-1)(1+r)^{-\frac{1}{2}}\log[2+|t-r|]\iint_{D''}\frac{1}{\tilde{w}(\lambda,s)}d\lambda ds\\ &+CH(t-r)\int_0^{t-r}\frac{1}{\tilde{w}(\lambda_-,s)}ds\int_{(\lambda_- -1)_+}^{\lambda_-}\frac{\lambda^{\frac{1}{2}}}{\{(\lambda+\lambda_-)(\lambda_+-\lambda)\}^{\frac{1}{2}}}\\ &\times\log\left[2+\frac{r\lambda}{(\lambda_- -\lambda)(\lambda_+ +\lambda)}\right]d\lambda.\end{aligned}$$

Notice that

$$\int_{(\lambda_- -1)_+}^{\lambda_-}\frac{\lambda^{\frac{1}{2}}}{\{(\lambda+\lambda_-)(\lambda_+-\lambda)\}^{\frac{1}{2}}}\log\left[2+\frac{r\lambda}{(\lambda_- -\lambda)(\lambda_+ +\lambda)}\right]d\lambda \le C\frac{\log[2+|t-r|]}{(1+r)^{\frac{1}{2}}}.$$

(For the proof, see (5.73) in [2].) Therefore we have

$$(4.48)\qquad (1+r)^{\frac{1}{2}}P_2 \le C\log[2+|t-r|]\left\{\iint_{D''}\frac{1}{\tilde{w}(\lambda,s)}d\lambda ds+\int_0^{t-r}\frac{ds}{1+s}\right\},$$

which implies (4.47) for $j=2$. This completes the proof. □

LEMMA 4.3. *Let F^i satisfy* (1.12) *and u be smooth function satisfying* (4.1) *with $k=[(N+1)/2]$. If we set*

$$(4.49)\qquad \frac{1}{w(\lambda,s)}=\sum_{j,k=1}^{m}\frac{1}{(w_jw_k)(\lambda,s)}\qquad \textit{for } \lambda>0, s>0,$$

then we have

$$(4.50)\qquad M_{0,N}\le C_N[\partial u]^2_{[\frac{N+1}{2}],t}\|\partial u\|_{N+3,t}.$$

Moreover, if we set

$$(4.51)\qquad \frac{1}{w(\lambda,s)}=\sum_{j,k,l=1}^{m}\frac{\lambda^{\frac{1}{2}}}{(w_jw_kw_l)(\lambda,s)}\ \textit{for } \lambda>0, s>0,$$

then we have

$$(4.52)\qquad M_{0,[\frac{N+1}{2}]}\le C_N[\partial u]^3_{[\frac{N+1}{2}]+1,t}.$$

Here $M_{0,N}$ is defined in (4.20).

Proof. First, we shall show (4.52). Since (4.1) with $k=[(N+1)/2]$ implies

$$(4.53)\ \sum_{j=1}^{m}\sum_{|a|\le[(N+1)/2]}|\Gamma^a\partial u^j(y,s)|\le[\partial u]_{[\frac{N+1}{2}],T}<1\quad \text{for } 0\le s\le t<T, y\in\mathbf{R}^2,$$

by (1.12) we have for $|a|\le[(N+1)/2]$

$$|\Gamma^aF^i(y,s)|\le C\sum_{j,k,l=1}^{m}\frac{1}{(w_jw_kw_l)(\lambda,s)}[\partial u(s)]^3_{[\frac{N+1}{2}]+1}$$

with $\lambda = |y|$. By (4.51), we therefore get (4.52).

Second, we shall prove (4.50). It follows that for $|a| \le N$

$$|\Gamma^a F^i(y,s)| \le C \sum_{j,k,l=1}^m \frac{1}{(w_j w_k)(\lambda,s)} [\partial u(s)]^2_{[\frac{N+1}{2}]} \sum_{|b|\le|a|+1} |\Gamma^b \partial u^l(y,s)|.$$

We now use an imbedding theorem concerning the invariant norm

$$|x|^{\frac{1}{2}}|f(x)| \le \sum_{|a|\le 2} \|\Gamma^a f\|_{L^2} \quad \text{for } x \in \mathbf{R}^2. \tag{4.54}$$

(For the proof, see, e.g., Lemma 6 in [19].) Applying this and using (4.49), we obtain (4.50). The proof is complete. □

COROLLARY 4.1. *Let $u = (u^1, \dots, u^m)$ be the solution of* (1.1) *and* (1.2) *and let F^i satisfy* (1.12). *Let $0 \le \mu < 1/2$. Then we have for $(|x|,t) \in \Lambda_i$ with $t < T$*

$$(1+t+r)^\mu |\Gamma^a u^i(x,t)| \le C_N \left(\varepsilon + [\partial u]^2_{[\frac{N+1}{2}],t} \|\partial u\|_{N+3,t} \right) \quad \textit{for } |a| \le N, \tag{4.55}$$

$$(1+t+r)^\mu |\Gamma^a u^i(x,t)| \le C_N \left(\varepsilon + [\partial u]^3_{[\frac{N+1}{2}],t} \right) \quad \textit{for } |a| \le [(N+1)/2], \tag{4.56}$$

provided (4.1) *with $k = [(N+1)/2]$ holds.*

Proof. Using the decomposition (4.5) with $b = 0$ and the estimates (4.6) and (4.45), we have

$$(1+t+r)^\mu |\Gamma^a u^i(x,t)| \le M_N \varepsilon + C_N M_{0,|a|}, \tag{4.57}$$

where $M_{0,|a|}$ is defined in (4.20), if $w(r,t)$ satisfies (4.19). Note that both (4.51) and (4.49) satisfy (4.19). Applying Lemma 4.3, we obtain (4.55) and (4.56). This completes the proof. □

End of the proof of Proposition 4.1. First we shall show (4.2). Using the decomposition (4.5) with $|b| = 1$ and the estimates (4.6) and (4.20), we have

$$w_i(r,t)|\Gamma^a \partial u^i(x,t)| \le M_N \varepsilon + C_N M_{0,|a|+1} \quad \text{for } (|x|,t) \in \tilde{\Lambda}_i^c \text{ with } t < T \tag{4.58}$$

if $w(r,t)$ satisfies (4.19). Using (4.50) with N replaced by $N+1$, we obtain (4.2).

Next we shall show (4.3). Similarly, it follows from (4.6) and (4.22) that for $|a| \le N$

$$w_i(r,t)|\Gamma^a \partial u^i(x,t)| \le M_N \varepsilon + C_N (M_{1,N+1} + M_{2,N+1} + M_{3,N+1}) \tag{4.59}$$

if $\eta_i(r,t)$ $(i = 1,2)$ satisfies (4.21).

First, we shall show

$$M_{1,N+1} \le C\left(\varepsilon + [\partial u]^2_{[\frac{N+4}{2}],t} \|\partial u\|_{N+6,t} \right). \tag{4.60}$$

If we set

$$\frac{1}{\eta_1(\lambda,s)} = \sum_{j,k,l=1}^m \frac{1}{(w_j w_k w_l)(\lambda,s)} + \sum_{(j,k)\ne(i,i)} \frac{1}{(w_j w_k)(\lambda,s)} + \frac{1-\tilde{\chi}(\lambda,s)}{\{w_i(\lambda,s)\}^2}, \tag{4.61}$$

then $\eta_1(r,t)$ satisfies the first condition in (4.21). Here $\tilde{\chi}$ is the characteristic function of $\tilde{\Lambda}_i$. In what follows, we always assume $|a| \leq N+1$. By (1.12) and (4.1) with $k = [(N+2)/2]$, we have

$$(4.62)\ |\Gamma^a G^i(y,s)| \leq C \sum_{j,k,l=1}^{m} \frac{1}{(w_j w_k w_l)(\lambda,s)} [\partial u(s)]^3_{\left[\frac{|a|+1}{2}\right]} \sum_{|b|\leq|a|+1} |\Gamma^b \partial u^l(y,s)|.$$

Using (4.54), we get

$$|\lambda^{\frac{1}{2}} \eta_1(\lambda,s) \Gamma^a G^i(y,s)| \leq C [\partial u(s)]^2_{\left[\frac{N+2}{2}\right]} \|\partial u(s)\|_{N+4}. \tag{4.63}$$

As for the resonance-form R^i, we find from (1.13) that there is at least one index among j, k, and l which does not coincide with i. Therefore, by (1.12) we have

$$\begin{aligned} |\Gamma^a R^i(y,s)| \leq{}& C \sum_{(j,k)\neq(i,i)} \sum_{l=1}^{m} \frac{1}{(w_j w_k)(\lambda,s)} [\partial u(s)]^2_{\left[\frac{|a|+1}{2}\right]} \sum_{|b|\leq|a|+1} |\Gamma^b \partial u^l(y,s)| \\ &+ C \sum_{\substack{(j,k)=(i,i)\\ l\neq i}} \frac{1-\tilde{\chi}(\lambda,s)}{\{w_i(\lambda,s)\}^2} [\partial u(s)]^2_{\left[\frac{|a|+1}{2}\right]} \sum_{|b|\leq|a|+1} |\Gamma^b \partial u^l(y,s)| \\ &+ C \sum_{\substack{(j,k)=(i,i)\\ l\neq i}} \frac{1}{(w_i w_l)(\lambda,s)} \frac{[\partial u(s)]^2_{\left[\frac{|a|+1}{2}\right]}}{w_i(\lambda,s)} \sum_{|b|\leq|a|+1} |(\tilde{\chi} w_l)(\lambda,s) \Gamma^b \partial u^l(y,s)|. \end{aligned} \tag{4.64}$$

By (4.61) and (4.54), we find that the first and second terms are dominated by

$$C\lambda^{-\frac{1}{2}} \{\eta_1(\lambda,s)\}^{-1} [\partial u(s)]^2_{\left[\frac{N+2}{2}\right]} \|\partial u(s)\|_{N+4}. \tag{4.65}$$

On the other hand, by (4.61), (4.1) with $k = [(N+2)/2]$, and $w_i(\lambda,s) \geq \lambda^{1/2}$, the third term is dominated by

$$C\lambda^{-\frac{1}{2}} \{\eta_1(\lambda,s)\}^{-1} \sum_{|b|\leq N+2} |(\tilde{\chi} w_l)(\lambda,s) \Gamma^b \partial u^l(y,s)|. \tag{4.66}$$

Moreover, since $\tilde{\Lambda}_i \subset \tilde{\Lambda}_l^c$ by (2.4), we get from (4.2),

$$|(\tilde{\chi} w_l)(\lambda,s) \Gamma^b \partial u^l(y,s)| \leq C_N \left(\varepsilon + [\partial u]^2_{\left[\frac{N+4}{2}\right],t} \|\partial u\|_{N+6,t} \right)$$

for $|b| \leq N+2$. We thus find that the third term is dominated by

$$C\lambda^{-\frac{1}{2}} \{\eta_1(\lambda,s)\}^{-1} \left(\varepsilon + [\partial u]^2_{\left[\frac{N+4}{2}\right],t} \|\partial u\|_{N+6,t} \right); \tag{4.67}$$

hence, together with (4.65), we get

$$|\lambda^{\frac{1}{2}} \eta_1(\lambda,s) \Gamma^a R^i(y,s)| \leq C \left(\varepsilon + [\partial u]^2_{\left[\frac{N+4}{2}\right],t} \|\partial u\|_{N+6,t} \right). \tag{4.68}$$

Combining (4.63) and (4.68), we finally get (4.60).

Second, we consider $M_{2,N+1}$. Taking $\eta_2(r,t) = w_i(r,t)^2$, we easily see that $\eta_2(r,t)$ satisfies the second condition of (4.21) and that

$$M_{2,N+1} \leq C[\partial u]^2_{[\frac{N+2}{2}],t} \|\partial u\|_{N+4,t}. \tag{4.69}$$

Third, we consider $M_{3,N+1}$ by taking $\eta_2(r,t) = w_i(r,t)^2$. By (3.1) we have

$$(1+s)^{\frac{1}{2}} |\Gamma^a N^i(y,s)| \leq C(\Phi^i_a + (1+s)^{-\frac{1}{2}} \Theta^i_a)$$

for $(|y|, s) \in \Lambda_i$. Therefore, we obtain

$$\begin{aligned} |\lambda^{\frac{1}{2}} \eta_2(\lambda,s)(1+s)^{\frac{1}{2}} \Gamma^a N^i(y,s)| &\leq C \sum_{|b+c+d| \leq |a|+1} |\lambda^{\frac{1}{2}} \eta_2(\lambda,s) |\partial \Gamma^b u^i \| \partial \Gamma^c u^i \| \partial \Gamma^d u^i \| \\ &+C \sum_{\substack{|b+c+d| \leq |a|+2 \\ |b|,|c|,|d| \leq |a|+1}} |\lambda^{\frac{1}{2}} \eta_2(\lambda,s)(1+s)^{-\frac{1}{2}} |\Gamma^b u^i \| \partial \Gamma^c u^i \| \partial \Gamma^d u^i \|. \end{aligned} \tag{4.70}$$

We easily see that the first term is dominated by $C[\partial u(s)]^2_{[(N+2)/2]} \|\partial u(s)\|_{N+4}$.

To treat the second term, we divide the argument into two cases. First we assume $|b| \geq [(N+2)/2]$. Since $1+\lambda$ is equivalent to $1+s$ for $(\lambda, s) \in \Lambda_i$ by (2.9), we have

$$\begin{aligned} &|\lambda^{\frac{1}{2}} \eta_2(\lambda,s)(1+s)^{-\frac{1}{2}} |\Gamma^b u^i \| \partial \Gamma^c u^i \| \partial \Gamma^d u^i \| \\ &\leq C[\partial u(s)]^2_{[\frac{N+2}{2}]} |\Gamma^b u^i(y,s)| \\ &\leq C\left(M_N \varepsilon + [\partial u]^2_{[\frac{N+4}{2}],s} \|\partial u\|_{N+6,s} \right), \end{aligned}$$

where we have used (4.1) with $k = [(N+4)/2]$ and (4.55) with $\mu = 0$ and N replaced by $N+3$.

Next we assume $|b| \leq [(N+2)/2]$. In this case, we have

$$\begin{aligned} &|\lambda^{\frac{1}{2}} \eta_2(\lambda,s)(1+s)^{-\frac{1}{2}} |\Gamma^b u^i \| \partial \Gamma^c u^i \| \partial \Gamma^d u^i \| \\ &\leq C \|\partial u(s)\|_{N+4} [\partial u(s)]_{[\frac{N+2}{2}]} |w_i(\lambda,s)(1+s)^{-\frac{1}{2}} |\Gamma^b u^i(\lambda,s)\| \\ &\leq C(1+s)^{\frac{1}{4}-\mu} \|\partial u(s)\|_{N+4} \left(\varepsilon + [\partial u]^3_{[\frac{N+2}{2}]+1,s} \right), \end{aligned}$$

where we have used (4.54), (2.8), (4.1) with $k = [(N+4)/2]$, and (4.56). Taking μ such that $\mu > 1/4$, we obtain

$$M_{3,N+1} \leq C\left((1 + \|\partial u\|_{N+4,t})\varepsilon + [\partial u]^2_{[\frac{N+4}{2}],t} \|\partial u\|_{N+6,t} \right). \tag{4.71}$$

Combining (4.60), (4.69), and (4.71) with (4.59), we obtain (4.3). This completes the proof. □

5. Proof of Theorem 1.1. By the existence and the uniqueness of the local smooth solution of (1.1) and (1.2) (see, e.g., S. Klainerman [14]), it is enough to establish a uniform a priori estimate of $[\partial u(t)]_N$ for some large integer N. To deal with the L^2-norm in the right-hand side of (4.3), we need the following.

Proposition 5.1. *Let $u^i \in C^\infty(\mathbf{R}^2 \times [0,T))$ be a solution of* (1.1) *and* (1.2). *Suppose that* (1.5) *holds. Then there exists a sufficiently small $\delta_1 > 0$ independent* *of T and a constant $C_N > 0$ independent of T and δ_1 such that the following energy estimate holds for $0 \le t < T$:*

$$\|\partial u(t)\|_N \le C_N \|\partial u(0)\|_N (1+t)^{C_N [\partial u]^2_{[\frac{N+1}{2}],t}}, \tag{5.1}$$

provided (4.1) *with $k = [(N+1)/2]$ holds.*

Proposition 5.2. *Let $u^i \in C^\infty(\mathbf{R}^2 \times [0,T))$ be a solution of* (1.1) *and* (1.2). *Also let $0 < \delta_1 < 1$ in* (4.1). *Suppose that* (1.11) *holds. Then there exists a constant $C_N > 0$ independent of T and δ_1 such that the following energy estimate holds for $0 \le t < T$:*

$$\begin{aligned} \|\partial u(t)\|_N^2 \le C_N^2 \Big\{ & \|\partial u(0)\|_N^2 \\ & + \int_0^t (1+s)^{-\frac{5}{4}} ([\partial u(s)]^2_{N+1} + \langle u(s)\rangle^2_{N+1}) \|\partial u(s)\|^2_{N+1} ds \Big\}, \end{aligned} \tag{5.2}$$

provided (4.1) *with $k = [(N+1)/2]$ holds. Here we have set*

$$\langle u(s)\rangle_k = \sum_{i=1}^m \sum_{|a|\le k} \sup_{\{x\in \mathbf{R}^2 : (x,s)\in \Lambda_i\}} |\Gamma^a u^i(x,s)|.$$

Proof of Proposition 5.1. If we set

$$L_i v = \Box_i v^i - \sum_{l=1}^m \sum_{\gamma,\delta=0}^2 H_{il}^{\gamma\delta}(\partial u)\partial_\gamma \partial_\delta v^l - K_i(\partial u) \quad \text{for} \quad v = (v^1, \dots, v^m), \tag{5.3}$$

we have an identity

$$\begin{aligned} \frac{d}{dt}\int_{\mathbf{R}^2} \Big\{ & (\partial_t v^i)^2 + c_i^2 |\nabla v^i|^2 - \sum_{l=1}^m H_{il}^{00}(\partial u)\partial_t v^i \partial_t v^l \\ & + \sum_{p,q=1}^2 H_{il}^{pq}(\partial u)\partial_p v^i \partial_q v^q \Big\} dx = \int_{\mathbf{R}^2} J_i(v) dx, \end{aligned} \tag{5.4}$$

where

$$\begin{aligned} J_i(v) = {} & 2L_i v \partial_t v^i - \sum_{l=1}^m (\partial_t H_{il}^{00}(\partial u))\partial_t v^i \partial_t v^l + 2\sum_{l=1}^m \sum_{p=1}^2 (\partial_p H_{il}^{p0}(\partial u))\partial_t v^i \partial_t v^l \\ & -2 \sum_{l=1}^m \sum_{p,q=1}^2 (\partial_p H_{il}^{pq}(\partial u))\partial_q v^i \partial_t v^l \\ & + \sum_{l=1}^m \sum_{p,q=1}^2 (\partial_t H_{il}^{pq}(\partial u))\partial_p v^i \partial_q v^l + 2K_i(\partial u)\partial_t v^i. \end{aligned}$$

Here we have used (1.5) and the divergence theorem. By (1.12), we have

$$|H_{il}^{\gamma\delta}(\partial u)| < \frac{1}{2m}\min\{1, c_m^2\}$$

if we take δ_1 in (4.1) to be sufficiently small. Therefore, (5.4) yields

$$\|\partial v(t)\|_0^2 \leq C\left(\|\partial v(0)\|_0^2 + \sum_{i=1}^{m}\int_0^t ds\int_{\mathbf{R}^2}|J_i(v)|dx\right). \tag{5.5}$$

Hence, if we take $v = \Gamma^a u (|a| \leq N)$ in (5.5), we have

$$\|\partial u(t)\|_N^2 \leq C\left(\|\partial u(0)\|_N^2 + \sum_{i=1}^{m}\sum_{|a|\leq N}\int_0^t ds\int_{\mathbf{R}^2}|J_i(\Gamma^a u)|dx\right). \tag{5.6}$$

Furthermore, it follows from (1.12), (4.1), and the Leibniz rule that

$$\int_{\mathbf{R}^2}|J_i(\Gamma^a u)|dx \leq C|\partial u(s)|^2_{[\frac{N+1}{2}]}\|\partial u(s)\|_N^2. \tag{5.7}$$

Thus, combining (5.6) and (5.7) and using Gronwall's inequality, we have

$$\|\partial u(t)\|_N \leq C_N\|\partial u(0)\|_N \exp\left(\int_0^t C_N|\partial u(s)|^2_{[\frac{N+1}{2}]}ds\right), \tag{5.8}$$

which yields (5.1), due to (2.5). This completes the proof. □

Proof of Proposition 5.2. Multiplying $\partial_t\Gamma^a u^i$ by (3.9) and integrating it over $\mathbf{R}^2 \times [0,t]$, we have

$$\|\partial u^i(t)\|_N^2 \leq \|\partial u^i(0)\|_N^2 + C_N\sum_{|b|\leq|a|\leq N}\int_0^t\int_{\mathbf{R}^2}|\Gamma^b(F^i(\partial u, \partial^2 u))\partial_t\Gamma^a u^i|dxds. \tag{5.9}$$

We divide the function F^i into three parts: G^i, R^i, and N^i as in (1.12).

First, we derive the estimate for the higher-order term G^i. Using (2.5) and (4.1) with $k = [(N+1)/2]$, we have

$$|\Gamma^b G^i(x,s)| \leq C_N(1+s)^{-\frac{3}{2}}[\partial u(s)]^3_{[\frac{N+1}{2}]}\sum_{j=1}^{m}\sum_{|c|\leq|b|+1}|\partial\Gamma^c u^j(x,s)|, \tag{5.10}$$

which yields

$$\int_{\mathbf{R}^2}|\Gamma^b(G^i(x,s))\partial_t\Gamma^a u^i|dx \leq C_N(1+s)^{-\frac{3}{2}}[\partial u(s)]^3_{[\frac{N+1}{2}]}\|\partial u(s)\|_{N+1}^2. \tag{5.11}$$

Second, we consider the resonance-form R^i. Without loss of generality, we may assume $l \neq i$ by (1.13). We now use the "resonance" property by the aid of (2.5), (2.6), and (2.4), namely,

$$\frac{1}{(w_l w_i)(|x|,s)} \leq \frac{C}{(1+s)^{\frac{5}{4}}}. \tag{5.12}$$

Using this estimate, we get

$$\begin{aligned}|\Gamma^b(R^i(x,s))\partial_t\Gamma^a u^i| &\leq C_N\sum_{j,k,=1}^{m}\sum_{l\neq i}\sum_{|c+d+e|\leq|b|+1}|\Gamma^c(\partial u^j)\Gamma^d(\partial u^k)\Gamma^e(\partial u^l)\partial_t\Gamma^a u^i|\\ &\leq C_N\sum_{j,k=1}^{m}\sum_{|c+d|\leq|b|+1}(1+s)^{-\frac{5}{4}}[\partial u(s)]^2_{N+1}|\Gamma^c(\partial u^j)\Gamma^d(\partial u^k)|,\end{aligned}$$

which yields

$$\int_{\mathbf{R}^2} |\Gamma^b(R^i(x,s))\partial_t\Gamma^a u^i| dx \le C_N(1+s)^{-\frac{5}{4}}[\partial u(s)]^2_{N+1}\|\partial u(s)\|^2_{N+1}. \tag{5.13}$$

Finally, we treat the null-form N^i. When $(x,s) \in \Lambda_i^c$, we find from (2.7) that

$$|\Gamma^b(N^i(x,s))| \le C_N(1+s)^{-\frac{3}{2}}[\partial u(s)]^2_{[\frac{N+1}{2}]} \sum_{|c|\le|b|+1} |\partial\Gamma^c u^i(x,s)|.$$

When $(x,s) \in \Lambda_i$, it follows from Proposition 3.1 and (2.5) that

$$\begin{aligned}|\Gamma^b(N^i(x,s))| &\le C_N((1+s)^{-\frac{1}{2}}\Phi^i_b + (1+s)^{-1}\Theta^i_b)\\ &\le C_N(1+s)^{-\frac{3}{2}}([\partial u(s)]^2_{N+1} + [\partial u(s)]_{N+1}\langle u(s)\rangle_{N+1}) \sum_{|c|\le|b|+1} |\partial\Gamma^c u^i(x,s)|.\end{aligned}$$

Therefore, we get

$$\begin{aligned}&\int_{\mathbf{R}^2} |\Gamma^b(N^i(x,s))\partial_t\Gamma^a u^i| dx\\ &\le \|\Gamma^b(N^i(s))\|_0 \|\partial u(s)\|_{N+1}\\ &\le C_N(1+s)^{-\frac{3}{2}}([\partial u(s)]^2_{N+1} + \langle u(s)\rangle^2_{N+1})\|\partial u(s)\|^2_{N+1}.\end{aligned} \tag{5.14}$$

Combining (5.11), (5.13), and (5.14) with (5.9), we obtain (5.2). The proof is complete. □

Corollary 5.1. *Let* $u^i \in C^\infty(\mathbf{R}^2 \times [0,T))$ *be a solution of* (1.1) *and* (1.2). *Suppose that* (1.5) *and* (1.11) *hold. Then there exist a sufficiently small* $\delta_1 > 0$ *independent of* T *and a constant* $C_N > 0$ *independent of* T *and* δ_1 *such that the following holds for* $0 \le t < T$:

$$\|\partial u(t)\|^2_{N+6} \le C_N^2\varepsilon^2 \left\{1 + \int_0^t (1+s)^{-\frac{5}{4}+4C_N[\partial u]^2_{[\frac{N+14}{2}],s}} ds\right\}, \tag{5.15}$$

provided (4.1) *with* $k = [(N+14)/2]$ *holds and* $0 < \varepsilon \le 1$.

Proof. It follows from (4.3) and (5.1) that for $0 \le s \le t$

$$[\partial u(s)]_{N+7} \le C_N(\varepsilon + (\varepsilon + \delta_1^2)\|\partial u\|_{N+13,s}) \tag{5.16}$$

and

$$\|\partial u\|_{N+13,s} \le C_N\varepsilon(1+s)^{C_N[\partial u]^2_{[\frac{N+14}{2}],s}} \tag{5.17}$$

because $\|\partial u(0)\|_{N+13} \le C_N\varepsilon$ for sufficiently small δ_1. Therefore, we have

$$[\partial u(s)]_{N+7} \le C_N(1+\varepsilon+\delta_1^2)\varepsilon(1+s)^{C_N[\partial u]^2_{[\frac{N+14}{2}],s}}. \tag{5.18}$$

Moreover, $\langle u(s)\rangle_{N+7}$ has the same estimate as $[\partial u(s)]_{N+7}$, because of (4.55). Now (5.15) follows from (5.2) and (5.18) together with (5.17). The proof is complete. □

End of the proof of Theorem 1.1. As we stated at the beginning of the present section, what we need to prove Theorem 1.1 is an a priori estimate for $[\partial u(t)]_N$. We fix an integer N satisfying $N \ge 13$, which guarantees $[(N+14)/2] \le N$. We take a positive constant B_N such that $B_N \geq 2\tilde{C}_N$ and $B_N \geq M_N$, where M_N is the constant in (4.6) and $\tilde{C}_N$ is the constant larger than C_N appearing in (4.3) and (5.15). We also take ε_1 such that

$$0 < \varepsilon_1 \leq 1 \quad \text{and} \quad 3B_N\varepsilon_1 \leq \delta_1, \tag{5.19}$$

where δ_1 is the smallest one taken in Proposition 4.1 and Corollary 5.1. Moreover, set

$$\begin{aligned} T_\varepsilon = \sup\{T > 0 :\ &\text{(1.1) and (1.2) have a solution } u^i \text{ in } C^\infty(\mathbf{R}^2 \times [0, T)) \\ &\text{and } [\partial u]_{N,T} \leq 3B_N\varepsilon \text{ holds}\}. \end{aligned} \tag{5.20}$$

We can see that $T_\varepsilon > 0$, because of the existence of a local solution, the continuity of $[\partial u]_{N,t}$, and (4.5). Then, for each ε satisfying $0 < \varepsilon \leq \varepsilon_1$, we have $u^i \in C^\infty(\mathbf{R}^2 \times [0, T_\varepsilon))$ and

$$[\partial u]_{[\frac{N+14}{2}],T_\varepsilon} \leq [\partial u]_{N,T_\varepsilon} \leq \delta_1,$$

which imply that (4.3) and (5.15) hold. In particular, we have for $0 \leq t < T_\varepsilon$

$$\|\partial u(t)\|_{N+6} \leq \tilde{C}_N\varepsilon \left\{1 + \int_0^t (1+s)^{-\frac{5}{4}+4\tilde{C}_N[\partial u]_{[\frac{N+14}{2}],s}} ds\right\}^{\frac{1}{2}}. \tag{5.21}$$

Now, we take ε_0 to be

$$0 < \varepsilon_0 \leq \varepsilon_1, \quad 3\tilde{C}_N\varepsilon_0 \leq 1, \quad \text{and} \quad 12\tilde{C}_N B_N\varepsilon_0 \leq \frac{1}{8}, \tag{5.22}$$

and fix an ε in $[0, \varepsilon_0)$ in the following. Then, by (5.21), (5.20), and (5.22), we have for $0 \leq t < T_\varepsilon$

$$\begin{aligned} \|\partial u(t)\|_{N+6} &\leq \tilde{C}_N\varepsilon \left(1 + \int_0^t (1+s)^{-\frac{9}{8}} ds\right)^{\frac{1}{2}} \\ &\leq 1. \end{aligned}$$

Substituting this into (4.3) and using (5.20), we have

$$[\partial u]_{N,T_\varepsilon} \leq \tilde{C}_N \left(2\varepsilon + 3B_N\varepsilon[\partial u]_{[\frac{N+4}{2}],T_\varepsilon}\right).$$

Hence, by $B_N \geq 2\tilde{C}_N$ and (5.22), we have

$$[\partial u]_{N,T_\varepsilon} \leq 2B_N\varepsilon. \tag{5.23}$$

By the blowup criterion (see, e.g., [22, Theorem 2.2, p. 31]), we see that if $T_\varepsilon < +\infty$, we must have $\lim_{t\to T_\varepsilon - 0}[\partial u]_{N,T} = 3B_N\varepsilon$, which contradicts (5.23). Therefore, we have $T_\varepsilon = +\infty$. This completes the proof of Theorem 1.1. □

Acknowledgments. The authors are grateful to Professor R. Agemi and Professor K. Kubota for their valuable comments.

REFERENCES

[1] R. Agemi, *Blow-up of solutions to nonlinear wave equations in two space dimensions*, Manuscripta Math., 73 (1991), pp. 153–162.

[2] R. Agemi and K. Yokoyama, *The null condition and global existence of solutions to systems of wave equations with different speeds*, in Advances in Nonlinear Partial Differential Equations and Stochastics, Ser. Adv. Math. Appl. Sci. 48, World Scientific, River Edge, NJ, 1998, pp. 43–86.

[3] S. Alinhac, *Temps de vie et comportement explosif des solutions déquations dondes quasilinéaires en dimension deux* I, Ann. Sci. École Norm. Sup., 28 (1995), pp. 225–251.

[4] D. Christodoulou, *Global solutions of nonlinear hyperbolic equations for small initial data*, Comm. Pure Appl. Math., 39 (1986), pp. 267–282.

[5] R. T. Glassey, *Existence in the large for $\Box u = F(u)$ in two space dimensions*, Math. Z., 178 (1981), pp. 233–261.

[6] P. Godin, *Lifespan of solutions of semilinear wave equations in two space dimensions*, Comm. Partial Differential Equations, 18 (1993), pp. 895–916.

[7] L. Hörmander, *The Lifespan of Classical Solutions of Nonlinear Hyperbolic Equations*, Lecture Notes in Math. 1256, Springer-Verlag, Berlin, New York, 1987.

[8] A. Hoshiga, *The initial value problems for quasi-linear wave equations in two space dimensions with small data*, Adv. Math. Sci. Appl., 5 (1995), pp. 67–89.

[9] A. Hoshiga, *The asymptotic behaviour of the radially symmetric solutions to quasilinear wave equations in two space dimensions*, Hokkaido Math. J., 24 (1995), pp. 575–615.

[10] F. John, *Blow-up of radial solutions of $u_{tt} = c^2(u_t)\Delta u$ in three space dimensions*, Mat. Apl. Comput., 4 (1985), pp. 3–18.

[11] F. John, *Existence for large times of strict solutions of nonlinear wave equations in three space dimensions for small initial data*, Comm. Pure. Appl. Math., 40 (1987), pp. 79–109.

[12] F. John and S. Klainerman, *Almost global existence to nonlinear wave equations in the three space dimensions*, Comm. Pure. Appl. Math., 37 (1984), pp. 443–455.

[13] S. Katayama, *Global existence for systems of nonlinear wave equations in two space dimensions,* II, Publ. Res. Inst. Math. Sci., 31 (1995), pp. 645–665.

[14] S. Klainerman, *Global existence for nonlinear wave equations*, Comm. Pure Appl. Math., 33 (1980), pp. 43–101.

[15] S. Klainerman, *Long time behavior of solutions to nonlinear wave equations*, in Proceedings of the International Congress of Mathematicians, Warsaw, 1982.

[16] S. Klainerman, *Uniform decay estimate and Lorentz invariance of the classical wave equation*, Comm. Pure Appl. Math., 38 (1985), pp. 321–332.

[17] S. Klainerman, *The Null Condition and Global Existence to Nonlinear Wave Equations*, Lectures in Appl. Math. 23, AMS, Providence, RI, 1986.

[18] S. Klainerman and T. Sideris, *On almost global existence for nonrelativistic wave equation in 3d*, Comm. Pure Appl. Math., 23 (1986), pp. 293–326.

[19] M. Kovalyov, *Long-time behaviour of solutions of a system of nonlinear wave equations*, Comm. Partial Differential Equations, 12 (1987), pp. 471–501.

[20] M. Kovalyov, *Resonance-type behaviour in a system of nonlinear wave equations*, J. Differential Equations, 77 (1989), pp. 73–83.

[21] M. Kovalyov and K. Tsutaya, *Erratum to the paper "Long-time behaviour of solutions of a system of nonlinear wave equations*," Comm. Partial Differential Equations, 12 (1987), pp. 471–501; Comm. Partial Differential Equations, 18 (1993), pp. 1971–1976.

[22] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws*, Appl. Math. Sci. 53, Springer-Verlag, New York, 1984.

[23] T. Sideris, *The null condition and global existence of nonlinear elastic waves*, Invent. Math., 123 (1996), pp. 323–342.

SIAM J. Math. Anal.
Vol. 31, No. 3, pp. 514–534

GLOBAL ASYMPTOTIC STABILITY OF TRAVELING WAVES IN DELAYED REACTION-DIFFUSION EQUATIONS*

HAL L. SMITH† AND XIAO-QIANG ZHAO†

Abstract. The existence and comparison theorem of solutions is first established for the quasi-monotone delayed reaction-diffusion equations on R by appealing to the theory of abstract functional differential equations. The global asymptotic stability, Liapunov stability, and uniqueness of traveling wave solutions are then proved by the elementary super- and subsolution comparison and squeezing methods.

Key words. delayed reaction-diffusion equations, comparison principle, super- and subsolutions, traveling waves

AMS subject classifications. 35R10, 35B40, 34K30, 58D25

PII. S0036141098346785

1. Introduction.

Traveling wave solutions have been widely studied for nonlinear reaction-diffusion equations modeling a variety of physical and biological phenomena (see, e.g., [3], [11], monograph [16], and references therein). More recently, Chen [1] studied the existence, uniqueness, and global asymptotic stability of traveling wave solutions in nonlocal evolution equations with bistable nonlinearities. A basic assumption in [1] is the comparison principle. Shen [12] investigated these problems for traveling wave solutions in temporally almost periodic reaction-diffusion equations with bistable nonlinearities. In [9], Ogiwara and Matano discussed the monotonicity and stability of pseudotraveling wave solutions in temporally or spatially periodic media as an application of their general theory on stable equilibria in order-preserving systems in the presence of symmetry.

Recently, great attention has also been paid to reaction-diffusion equations with time delays (see, e.g., [15], [7], [8], [4], monograph [17], and references therein). Most of the known results in this direction are about the existence, comparison, monotonicity, bifurcations, and asymptotic behavior of solutions to delayed reaction-diffusion equations on a bounded spatial domain. Schaaf [13] first studied traveling wave solutions for some delayed reaction-diffusion equations and, in particular, proved the existence of monotone traveling wave solutions and uniqueness of wave speeds for the delayed reaction-diffusion equations with quasi-monotone and bistable nonlinearities by a phase plane analysis method. In [18], Zou and Wu obtained the existence of traveling waves in some delayed reaction-diffusion systems via the monotone iteration method. As a consequence of the delayed reaction term, the study of uniqueness and global asymptotic stability of traveling wave solutions becomes relatively more difficult. This paper is devoted to the study of global asymptotic stability with phase shift, Liapunov stability, and uniqueness up to translation of traveling wave solutions in delayed reaction-diffusion equations with quasi-monotone and bistable nonlinearities. The first key point is to establish a refined comparison principle for this class of delayed reaction-diffusion equations defined on the whole real line R. We do this by appealing to the theory of abstract functional differential equations developed in [7]

*Received by the editors November 6, 1998; accepted for publication (in revised form) June 23, 1999; published electronically February 10, 2000.

http://www.siam.org/journals/sima/31-3/34678.html

†Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804 (halsmith@asu.edu, xzhao@math.mun.ca). The first author was supported by NSF grant DMS 9700910.

and properties of the analytic semigroup generated by the one-dimensional Laplacian operator on the Banach space of all bounded and uniformly continuous functions on R. In order to prove global asymptotic stability of monotone traveling wave solutions, we have borrowed a "squeezing" technique introduced in [1], which is similar in spirit to a "contracting rectangles" approach developed in [8] for quasi-monotone delayed reaction-diffusion systems on a bounded spatial domain. The Liapunov stability of monotone traveling waves and the uniqueness of traveling waves are then proved by using an elementary super- and subsolution comparison method and the established global asymptotic stability of monotone traveling waves.

We note that the recent publication [9] contains results related to ours on traveling wave solutions for delayed reaction-diffusion equations. These authors show that monotone traveling waves are *locally* asymptotically stable with phase shift. By contrast, our results allow replacing locally with *globally and exponentially* in their result and are based on more elementary methods.

The organization of this paper is as follows. In section 2, we establish an existence and comparison theorem for quasi-monotone delayed reaction-diffusion equations on R (Theorem 2.2). For use in the next section, we also prove three technical lemmas about the construction of super- and subsolutions and the derivative of profiles of monotone traveling wave solutions (Lemmas 2.3, 2.4, and 2.5). In section 3, we first prove two lemmas about the iteration and ultimate estimation of solutions (Lemmas 3.1 and 3.2); then we establish the global exponential stability with phase shift of monotone traveling wave solutions (Theorem 3.3), Liapunov stability, and uniqueness up to translation of traveling wave solutions (Theorem 3.4).

2. Existence and comparison of solutions.

Consider delayed reaction-diffusion equations

$$\begin{cases} \dfrac{\partial u}{\partial t} &= d\Delta u + f(u(x,t), u(x,t-\tau)), \quad x \in R, t > 0, \\ u(x,s) &= \varphi(x,s), \quad x \in R, s \in [-\tau, 0], \end{cases} \tag{2.1}$$

where $d > 0$, $\tau \geq 0$, and Δ is the Laplacian operator on R. We will impose the following conditions on $f(\cdot,\cdot)$.

(H1) $f \in C^1(I^2, R)$ for some open interval $I \subset R$ with $[0,1] \subset I$; $\partial_2 f(u,v) \geq 0$, for $(u,v) \in I^2$.

(H2) $f(0,0) = f(1,1) = 0$, $\partial_1 f(0,0) + \partial_2 f(0,0) < 0$, and $\partial_1 f(1,1) + \partial_2 f(1,1) < 0$.

Let $X = BUC(R,R)$ be the Banach space of all bounded and uniformly continuous functions from R into R with the usual supremum norm. Let $X^+ = \{\varphi \in X; \varphi(x) \geq 0, x \in R\}$. It is easy to see that X^+ is a closed cone of X and its induced partial ordering makes X into a Banach lattice. By [2, Theorem 1.5], it then follows that the X-realization $d\Delta_X$ of $d\Delta$ generates a strongly continuous analytic semigroup $T(t)$ on X and $T(t)X^+ \subset X^+$, $t \geq 0$. Moreover, by the explicit expression of solutions of the heat equation

$$\begin{cases} \dfrac{\partial u}{\partial t} &= d\Delta u, \quad x \in R, t > 0, \\ u(x,0) &= \varphi(x), \quad x \in R, \end{cases} \tag{2.2}$$

we have

$$T(t)\varphi(x) = \frac{1}{\sqrt{4\pi dt}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-y)^2}{4dt}\right) \varphi(y)dy, \qquad x \in R, \quad t > 0, \quad \varphi_{(\cdot)} \in X. \tag{2.3}$$

Let $f_0(\cdot) : I \to R$ be defined by $f_0(u) = f(u, u)$, $u \in I$. By the continuity of f_0 and condition (H2), it then easily follows that there exist δ_0, a^-, $a^+ \in (0, 1)$ with $[-\delta_0, 1 + \delta_0] \subset I$ and $a^- \leq a^+$ such that $f_0(\cdot) : [-\delta_0, 1 + \delta_0] \to R$ satisfies

$$\begin{aligned} & f_0(0) = f_0(a^-) = f_0(a^+) = f_0(1) = 0, \\ & f_0(u) > 0 \quad \text{for } u \in [-\delta_0, 0) \cup (a^+, 1), \\ & \text{and } f_0(u) < 0 \quad \text{for } u \in (0, a^-) \cup (1, 1 + \delta_0]. \end{aligned} \tag{2.4}$$

Let $L_i = \max\{|\partial_i f(u, v)|; -\delta_0 \leq u, v \leq 1 + \delta_0\}$, $i = 1, 2$, and define

$$\theta(J, t) = \frac{1}{\sqrt{4\pi dt}} \exp\left(-L_1 t - \frac{(J + 1)^2}{4dt}\right), \quad J \geq 0,\ t > 0.$$

Clearly, $\theta \in C([0, \infty) \times (0, \infty), R)$.

Let $C = C([-\tau, 0], X)$ be the Banach space of continuous functions from $[-\tau, 0]$ into X with the supremum norm, and let $C^+ = \{\varphi \in C; \varphi(s) \in X^+, s \in [-\tau, 0]\}$. Then C^+ is a positive cone of C. For convenience, we identify an element $\varphi \in C$ as a function from $R \times [-\tau, 0]$ into R defined by $\varphi(x, s) = (\varphi(s))(x)$. For any continuous function $w(\cdot) : [-\tau, b) \to X$, $b > 0$, we define $w_t \in C$, $t \in [0, b)$, by $w_t(s) = w(t + s)$, $s \in [-\tau, 0]$. It is then easy to see that $t \mapsto w_t$ is a continuous function from $[0, b)$ to C. For any $\varphi \in [-\delta_0, 1 + \delta_0]_C = \{\varphi \in C; \varphi(x, s) \in [-\delta_0, 1 + \delta_0], x \in R, s \in [-\tau, 0]\}$, define $F(\varphi)(x) = f(\varphi(x, 0), \varphi(x, -\tau))$, $x \in R$. By the global Lipschitz continuity of $f(\cdot, \cdot)$ on $[-\delta_0, 1 + \delta_0]^2$, we can verify that $F(\varphi) \in X$ and $F : [-\delta_0, 1 + \delta_0]_C \to X$ is globally Lipschitz continuous.

Definition 2.1. *A continuous function $v : [-\tau, b) \to X$, $b > 0$, is called a supersolution (subsolution) of* (2.1) *on* $[0, b)$ *if*

$$v(t) \geq (\leq) T(t - s)v(s) + \int_s^t T(t - r)F(v_r)dr \tag{2.5}$$

for all $0 \leq s < t < b$. If v is both a supersolution and a subsolution on $[0, b)$*, then we call it a mild solution of* (2.1).

Remark 2.1. *Assume that there is a $v \in BUC(R \times [-\tau, b], R)$, $b > 0$, such that v is C^2 in $x \in R$, C^1 in $t \in (0, b)$, and*

$$\frac{\partial v}{\partial t} \geq (\leq) d\Delta v + f(v(x, t), v(x, t - \tau)), \quad x \in R, t \in (0, b). \tag{2.6}$$

Then, by the positivity of the linear semigroup $T(t) : X \to X$, it easily follows that (2.5) *holds. Hence v is a supersolution (subsolution) of* (2.1) *on* $[0, b)$.

Now we are in a position to establish the following existence and comparison theorem for (2.1).

Theorem 2.2. *Assume that* (H1) *and* (H2) *hold. Then for any $\varphi \in [-\delta_0, 1+\delta_0]_C$,* (2.1) *has a unique mild solution $u(x, t, \varphi)$ on $[0, \infty)$ and $u(x, t, \varphi)$ is a classical solution to* (2.1) *for $(x, t) \in R \times (\tau, \infty)$. Furthermore, for any pair of supersolution $u(x, t)$ and subsolution $w(x, t)$ of* (2.1) *on $[0, \infty)$ with $-\delta_0 \leq u(x, t)$, $w(x, t) \leq 1 + \delta_0$, $x \in R$, $t \in [-\tau, \infty)$, and $u(x, s) \geq w(x, s)$, $x \in R$, $s \in [-\tau, 0]$, there holds $u(x, t) \geq w(x, t)$ for $x \in R$, $t \geq 0$, and*

$$u(x, t) - w(x, t) \geq \theta(J, t - t_0) \int_z^{z+1} (u(y, t_0) - w(y, t_0))dy$$

for any $J \geq 0$*,* x *and* $z \in R$ *with* $|x - z| \leq J$*, and* $t > t_0 \geq 0$*.*

Proof. Under an abstract setting in [7], a mild solution of (2.1) is a solution to its associated integral equation

$$\begin{cases} u(t) = T(t)\varphi(0) + \int_0^t T(t-r)F(u_r)dr, & t > 0, \\ u_0 = \varphi \in [-\delta_0, 1+\delta_0]_C. \end{cases} \tag{2.7}$$

By the choice of δ_0 in (2.4), we have $f(1+\delta_0, 1+\delta_0) < 0$ and $f(-\delta_0, -\delta_0) > 0$. Clearly, $v^+ = 1 + \delta_0$ and $v^- = -\delta_0$ are an ordered pair of super- and subsolutions of (2.1) on $[0, \infty)$. As aforementioned, $F : [-\delta_0, 1+\delta_0]_C \to X$ is globally Lipschitz continuous. We further claim that F is quasi-monotone on $[-\delta_0, 1+\delta_0]_C$ in the sense that

$$\lim_{h \to 0^+} \frac{1}{h} \operatorname{dist}(\psi(0) - \varphi(0) + h[F(\psi) - F(\varphi)]; X^+) = 0 \tag{2.8}$$

$$\text{for all } \psi, \varphi \in [-\delta_0, 1+\delta_0]_C \text{ with } \psi \geq \varphi.$$

Indeed it follows from condition (H1) that

$$\begin{aligned} F(\psi) - F(\varphi) &= f(\psi(\cdot, 0), \psi(\cdot, -\tau)) - f(\varphi(\cdot, 0), \varphi(\cdot, -\tau)) \\ &\geq f(\psi(\cdot, 0), \varphi(\cdot, -\tau)) - f(\varphi(\cdot, 0), \varphi(\cdot, -\tau)) \\ &\geq -L_1(\psi(0) - \varphi(0)) \quad \text{in } X, \end{aligned} \tag{2.9}$$

and hence, for any $h > 0$ satisfying $hL_1 < 1$,

$$\psi(0) - \varphi(0) + h[F(\psi) - F(\varphi)] \geq (1 - L_1 h)(\psi(0) - \varphi(0)) \geq 0 \quad \text{in } X.$$

Then the existence and uniqueness of $u(x, t, \varphi)$ follows from [7, Corollary 5] with $S(t, s) = T(t, s) = T(t - s)$, $t \geq s \geq 0$, and $B(t, \varphi) \equiv F(\varphi)$. Moreover, by a semigroup theory argument given in the proof of [7, Theorem 1], it follows that $u(x, t, \varphi)$ is a classical solution for $t > \tau$.

For simplicity, let $\psi(x, s) = u(x, s)$, $\varphi(x, s) = w(x, s)$, $x \in R$, $s \in [-\tau, 0]$. Then $\psi, \varphi \in [-\delta_0, 1+\delta_0]_C$ with $\psi \geq \varphi$ in C. Again by [7, Corollary 5], we have

$$1 + \delta_0 \geq u(x, t, \psi) \geq u(x, t, \varphi) \geq -\delta_0, \quad x \in R, t \geq 0. \tag{2.10}$$

By applying [7, Corollary 5] with $v^+(x, t) = 1 + \delta_0$ and $v^-(x, t) = w(x, t)$, $v^+(x, t) = u(x, t)$ and $v^-(x, t) = -\delta_0$, respectively, we get

$$w(x, t) \leq u(x, t, \varphi) \leq 1 + \delta_0, \quad x \in R, t \geq 0 \tag{2.11}$$

and

$$-\delta_0 \leq u(x, t, \psi) \leq u(x, t), \quad x \in R, t \geq 0. \tag{2.12}$$

Combining (2.10)–(2.12), we have $u(x, t) \geq w(x, t)$ for all $x \in R$ and $t \geq 0$.

It remains to prove the last inequality in the theorem. Let $v(x, t) = u(x, t) - w(x, t)$, $x \in R$, $t \in [-\tau, \infty)$. Then $v(x, t) \geq 0$, $x \in R$, $t \in [-\tau, \infty)$. Clearly, $u_t, w_t \in [-\delta_0, 1+\delta_0]_C$ with $u_t \geq w_t$ in C for all $t \geq 0$. For any given $t_0 \geq 0$, by Definition 2.1 and (2.9), it then follows that, for all $t \geq t_0$,

$$\begin{aligned} v(t) &\geq T(t - t_0)v(t_0) + \int_{t_0}^t T(t - r)(F(u_r) - F(w_r))dr \\ &\geq T(t - t_0)v(t_0) - L_1 \int_{t_0}^t T(t - r)v(r)dr. \end{aligned} \tag{2.13}$$

Let

$$z(t) = e^{-L_1(t-t_0)}T(t-t_0)v(t_0), \quad t \geq t_0.$$

Then $z(t)$ satisfies

$$z(t) = T(t-t_0)z(t_0) - L_1 \int_{t_0}^{t} T(t-r)z(r)dr, \quad t \geq t_0. \tag{2.14}$$

By [7, Proposition 3] with $v^- \equiv z(t)$, $v^+ = +\infty$, $S(t,s) = S^-(t,s) = T(t,s) = T(t-s)$, $t \geq s \geq 0$, and $B(t,\varphi) \equiv B^-(t,\varphi) \equiv -L_1\varphi(0)$, we have $v(t) \geq z(t)$ for all $t \geq t_0$. Thus it follows that

$$u(t) - w(t) \geq e^{-L_1(t-t_0)}T(t-t_0)(u(t_0) - w(t_0)), \quad t \geq t_0. \tag{2.15}$$

Combining (2.3), (2.15), and the definition of $\theta \in C([0,\infty) \times (0,\infty), R)$, we then have

$$u(x,t) - w(x,t) \geq \theta(J, t-t_0) \int_{z}^{z+1} (u(y,t_0) - w(y,t_0))dy \tag{2.16}$$

for all $x \in R$ with $|x - z| \leq J$ and $t > t_0 \geq 0$.

This completes the proof. □

For delayed reaction-diffusion equation (2.1), we are interested in its traveling wave solutions which connect two equilibria 0 and 1. Without loss of generality, throughout this paper, by a traveling wave we refer to a solution of the form of $u(x,t) = U(x-ct)$, $x \in R$, $t \in R$, with the property that

$$\lim_{\xi \to \infty} U(\xi) = 1, \qquad \lim_{\xi \to -\infty} U(\xi) = 0, \tag{2.17}$$

where $U(\xi)$ is a function on R and c is a constant real number. As usual, $|c|$ is called the wave speed and U the profile of the wave front. Moreover, we say a traveling wave $U(x-ct)$ is a monotone if $U(\cdot) : R \to R$ is a strictly increasing function.

LEMMA 2.3. *Assume that* (H1) *and* (H2) *hold and let* $U(x-ct)$ *be a monotone traveling wave solution. Then there exist three positive numbers* β_0 *(which is independent of* U*),* σ_0*, and* $\bar{\delta}$ *such that for any* $\delta \in (0, \bar{\delta}]$ *and every* $\xi_0 \in R$*, the functions* w^+ *and* w^- *defined by*

$$w^{\pm}(x,t) := U(x - ct + \xi_0 \pm \sigma_0\delta[1 - e^{-\beta_0 t}]) \pm \delta e^{-\beta_0 t}$$

are a supersolution and a subsolution of (2.1) *on* $[0,\infty)$*, respectively.*

Proof. Clearly, $0 < U(\xi) < 1$, and hence, $0 < U(x-ct) < 1$, $x \in R$, $t \in R$. By Theorem 2.2 and monotonicity of $U(\cdot)$, it follows that $U(\cdot) \in C^1(R)$ and $U'(\xi) > 0$, $\xi \in R$. Since

$$\lim_{(u,v,\beta)\to(0,0,0)} (\partial_1 f(u,v) + e^{\beta\tau}\partial_2 f(u,v) + \beta) = \partial_1 f(0,0) + \partial_2 f(0,0) < 0$$

and

$$\lim_{(u,v,\beta)\to(1,1,0)} (\partial_1 f(u,v) + e^{\beta\tau}\partial_2 f(u,v) + \beta) = \partial_1 f(1,1) + \partial_2 f(1,1) < 0,$$

we can fix $\beta_0 > 0$ and $\delta^* \in (0, \delta_0)$ such that

$$\partial_1 f(u,v) + e^{\beta_0\tau}\partial_2 f(u,v) < -\beta_0 \text{ for all } (u,v) \in [-\delta^*, \delta^*]^2 \cup [1-\delta^*, 1+\delta^*]^2. \tag{2.18}$$

Let $c_0 = \tau|c| + (e^{\beta_0\tau} - 1)$. By (2.17), there exists $M_0 = M_0(U, \beta_0, \delta^*) > 0$ such that

$$\begin{aligned} &U(\xi) \geq 1 - \frac{\delta^*}{2} \quad \text{for all } \xi \geq M_0 - c_0, \\ &U(\xi) \leq \frac{\delta^*}{2} \quad \text{for all } \xi \leq -M_0 + c_0. \end{aligned} \tag{2.19}$$

Take

$$\begin{aligned} c_1 = c_1(\beta_0, \delta^*) &= \max\{|\partial_1 f(u,v)|; (u,v) \in [-\delta^*, 1+\delta^*]^2\} \\ &+ e^{\beta_0\tau} \max\{|\partial_2 f(u,v)|; (u,v) \in [-\delta^*, 1+\delta^*]^2\}, \end{aligned}$$

$m_0 = m_0(U, \beta_0, \delta^*) = \min\{U'(\xi); |\xi| \leq M_0\} > 0$, and define

$$\sigma_0 = \sigma_0(U, \beta_0, \delta^*) = \frac{\beta_0 + c_1}{m_0\beta_0} > 0, \quad \bar{\delta} = \bar{\delta}(U, \beta_0, \delta^*) = \min\{\frac{1}{\sigma_0}, \frac{\delta^*}{2} \cdot e^{-\beta_0\tau}\} > 0.$$

We prove only $w^+(x,t)$ is a supersolution of (2.1) since the proof for $w^-(x,t)$ is analogous. By a translation, we can assume that $\xi_0 = 0$. For any given $\delta \in (0, \bar{\delta}]$, let $\xi(t) = x - ct + \sigma_0\delta[1 - e^{-\beta_0 t}]$. It then follows that $\xi(t) + c\tau \geq \xi(t-\tau)$ and $|\xi(t-\tau) - \xi(t)| = |c\tau + \sigma_0\delta e^{-\beta_0 t}[1 - e^{\beta_0\tau}]| \leq c_0$ for all $t \geq 0$. Since $U(x - ct)$ is a solution of (2.1), we have

$$dU''(\xi) + cU'(\xi) + f(U(\xi), U(\xi + c\tau)) = 0, \xi \in R. \tag{2.20}$$

Notice that $U'(\xi) > 0$ and $\partial_2 f(\cdot,\cdot) \geq 0$. It then follows that, for any $t \geq 0$,

$$\begin{aligned} &\frac{\partial w^+(x,t)}{\partial t} - d\Delta w^+(x,t) - f(w^+(x,t), w^+(x,t-\tau)) \\ &= U'(\xi(t))(-c + \sigma_0\delta\beta_0 e^{-\beta_0 t}) - \beta_0\delta e^{-\beta_0 t} \\ &\quad -dU''(\xi(t)) - f(U(\xi(t)) + \delta e^{-\beta_0 t}, U(\xi(t-\tau)) + \delta e^{-\beta_0(t-\tau)}) \\ &= (U'(\xi(t))\sigma_0\beta_0 - \beta_0)\delta e^{-\beta_0 t} + f(U(\xi(t)), U(\xi(t) + (c\tau))) \\ &\quad -f(U(\xi(t)) + \delta e^{-\beta_0 t}, U(\xi(t-\tau) + \delta e^{-\beta_0(t-\tau)}) \\ &\geq (U'(\xi(t))\sigma_0\beta_0 - \beta_0)\delta e^{-\beta_0 t} + f(U(\xi(t)), U(\xi(t-\tau))) \\ &\quad -f(U(\xi(t)) + \delta e^{-\beta_0 t}, U(\xi(t-\tau)) + \delta e^{-\beta_0(t-\tau)}) \\ &= \delta e^{-\beta_0 t}\Big[U'(\xi(t))\sigma_0\beta_0 - \beta_0 - \int_0^1 \big(\partial_1 f(U(\xi(t)) + s\delta e^{-\beta_0 t}, U(\xi(t-\tau)) \\ &\quad + s\delta e^{-\beta_0(t-\tau)}) + e^{\beta_0\tau}\partial_2 f(U(\xi(t)) + s\delta e^{-\beta_0 t}, U(\xi(t-\tau)) + s\delta e^{-\beta_0(t-\tau)})\big)ds\Big]. \end{aligned} \tag{2.21}$$

We distinguish among three cases.

Case (i) $|\xi(t)| \leq M_0$: By the choice of M_0 and c_1, the absolute value of the integral in (2.21) is less than or equal to c_1. Then, by the choice of σ_0, we have

$$\begin{aligned} &\frac{\partial w^+(x,t)}{\partial t} - d\Delta w^+(x,t) - f(w^+(x,t), w^+(x,t-\tau)) \\ &\geq [m_0\sigma_0\beta_0 - \beta_0 - c_1] \cdot \delta e^{-\beta_0 t} = 0. \end{aligned} \tag{2.22}$$

Case (ii) $\xi(t) \geq M_0$: Clearly, $\xi(t-\tau) = \xi(t) + \xi(t-\tau) - \xi(t) \geq M_0 - c_0$. Then, by (2.19), $1 - \frac{\delta^*}{2} \leq U(\xi(t))$, $U(\xi(t-\tau)) \leq 1$. Thus we have

$$1 - \frac{\delta^*}{2} \leq U(\xi(t)) + \delta e^{-\beta_0 t} \leq 1 + \frac{\delta^*}{2}$$

and

$$1 - \frac{\delta^*}{2} \le U(\xi(t-\tau)) + \delta e^{-\beta_0(t-\tau)} \le 1 + \delta e^{\beta_0\tau} \cdot e^{-\beta_0 t} \le 1 + \frac{\delta^*}{2}.$$

Therefore, by (2.21) and (2.18), it follows that

$$\begin{aligned} &\frac{\partial w^+(x,t)}{\partial t} - d\Delta w^+(x,t) - f(w^+(x,t), w^+(x,t-\tau)) \\ &\ge [U'(\xi(t))\sigma_0\beta_0 - \beta_0 - (-\beta_0)]\delta e^{-\beta_0 t} \ge 0. \end{aligned} \tag{2.23}$$

Case (iii) $\xi(t) \le -M_0$: Clearly, $\xi(t-\tau) \le \xi(t) + \xi(t-\tau) - \xi(t) \le -M_0 + c_0$. Then, again by (2.19), $0 \le U(\xi(t)),\ U(\xi(t-\tau)) \le \frac{\delta^*}{2}$, and hence,

$$0 \le U(\xi(t)) + \delta e^{-\beta_0 t} \le \frac{\delta^*}{2} + \delta e^{-\beta_0 t} \le \delta^*$$

and

$$0 \le U(\xi(t-\tau)) + \delta e^{-\beta_0(t-\tau)} \le \frac{\delta^*}{2} + \delta e^{\beta_0\tau} \cdot e^{-\beta_0 t} \le \delta^*.$$

Again (2.23) follows from (2.21) and (2.18).
Combining cases (i)–(iii), we have

$$\frac{\partial w^+}{\partial t} - d\Delta w^+ - f(w^+(x,t), w^+(x,t-\tau)) \ge 0, \quad x \in R,\ t \ge 0. \tag{2.24}$$

This completes the proof. □

Let $\bar{\delta}_0 = \min\{\frac{a^-}{2}, \frac{1-a^+}{2}, \delta_0\}$, and let $\zeta(\cdot) \in C^\infty(R)$ be a fixed function with the following properties:

$$\begin{aligned} &\zeta(s) = 0 \quad \text{if } s \le 0; \qquad \zeta(s) = 1 \quad \text{if } s \ge 4; \\ &0 < \zeta'(s) < 1; \qquad |\zeta''(s)| \le 1 \quad \text{if } s \in (0,4). \end{aligned} \tag{2.25}$$

Then we have the following result.

Lemma 2.4. *Assume that* (H1) *and* (H2) *hold. Then, for any* $\delta \in (0, \bar{\delta}_0]$, *there exist two positive numbers* $\epsilon = \epsilon(\delta)$ *and* $C = C(\delta)$ *such that, for every* $\xi \in R$, *the functions* v^+ *and* v^- *defined by*

$$\begin{aligned} v^+(x,t) &:= (1+\delta) - [1 - (a^- - 2\delta)e^{-\epsilon t}]\zeta(-\epsilon(x - \xi + Ct)), \\ v^-(x,t) &:= -\delta + [1 - (1 - a^+ - 2\delta)e^{-\epsilon t}]\zeta(\epsilon(x - \xi - Ct)) \end{aligned}$$

are a supersolution and a subsolution of (2.1) *on* $[0,\infty)$, *respectively.*

Proof. By a translation, we can assume $\xi = 0$. Given $\delta \in (0, \bar{\delta}_0]$, we define

$$\begin{aligned} m_1 &= m_1(\delta) = \max\{\partial_2 f(u,v); (u,v) \in [-\delta, 1-\delta]^2\} \ge 0, \\ m_2 &= m_2(\delta) = \min\left\{\zeta'(s); \frac{\delta}{2} \le \zeta(s) \le 1 - \frac{\delta}{2}\right\} > 0. \end{aligned}$$

Then there exists an $\epsilon = \epsilon(\delta) > 0$ such that

$$(1 - a^+)e^{\epsilon\tau} < 1, \tag{2.26}$$

$$-\min\left\{f_0(u); u \in \left[-\delta, -\frac{\delta}{2}\right]\right\} + (\epsilon + \tau m_1 \epsilon + d\epsilon^2) < 0, \tag{2.27}$$

and

$$-\min\left\{f_0(u); u \in \left[a^+ + \frac{\delta}{2}, 1-\delta\right]\right\} + (\epsilon + \tau m_1 \epsilon + d\epsilon^2) < 0. \tag{2.28}$$

We further choose $C = C(\delta) > 0$ such that

$$-C\epsilon a^+ m_2 + \max\{|f_0(u)|; u \in [-\delta, 1-\delta]\} + (\epsilon + \tau m_1 \epsilon + d\epsilon^2) < 0. \tag{2.29}$$

By a direct computation and (2.26), it follows that for all $t \geq -\tau$,

$$\begin{aligned}\frac{\partial v^-(x,t)}{\partial t} &= -C\epsilon[1-(1-a^+-2\delta)e^{-\epsilon t}]\zeta'(\epsilon(x-Ct)) \\ &\quad +\epsilon(1-a^+-2\delta)\cdot\zeta(\epsilon(x-Ct))e^{-\epsilon t} \\ &\leq -C\epsilon[1-(1-a^+)e^{\epsilon\tau}]\zeta'(\epsilon(x-Ct)) + \epsilon(1-a^+)\zeta(\epsilon(x-Ct))e^{\epsilon\tau} \\ &\leq \epsilon. \end{aligned} \tag{2.30}$$

It is easy to see that $v^-(x,t) \in [-\delta, 1-\delta]$ for all $x \in R$ and $t \geq -\tau$. Therefore we have that for all $t \geq 0$,

$$\begin{aligned} f(v^-(x,t), v^-(x,t-\tau)) &= f_0(v^-(x,t)) + [f(v^-(x,t), v^-(x,t-\tau)) \\ &\quad -f(v^-(x,t), v^-(x,t))] \\ &= f_0(v^-(x,t)) + \partial_2 f(v^-(x,t), v^*(x,t))(v^-(x,t-\tau) \\ &\quad -v^-(x,t)) \\ &= f_0(v^-(x,t)) + \partial_2 f_2(v^-(x,t), v^*(x,t))\frac{\partial v^-(x,t^*)}{\partial t}\cdot \\ &\quad ((t-\tau)-t) \\ &\geq f_0(v^-(x,t)) - \tau\epsilon m_1, \end{aligned} \tag{2.31}$$

where $v^*(x,t)$ is between $v^-(x,t)$ and $v^-(x,t-\tau)$, $t^* \in [t-\tau, t]$, and hence, $t^* \geq -\tau$. It then follows that

$$\begin{aligned} &\frac{\partial v^-(x,t)}{\partial t} - d\Delta v^-(x,t) - f(v^-(x,t), v^-(x,t-\tau)) \\ &\quad = -C\epsilon[1-(1-a^+-2\delta)e^{-\epsilon t}]\zeta'(\epsilon(x-Ct) + \epsilon(1-a^+-2\delta)\zeta(\epsilon(x-Ct))\cdot \\ &\quad e^{-\epsilon t} - d[1-(1-a^+-2\delta)e^{-\epsilon t}]\zeta''(\epsilon(x-Ct))\epsilon^2 - f(v^-(x,t), v^-(x,t-\tau)) \\ &\quad \leq -C\epsilon a^+\zeta'(\epsilon(x-Ct)) + \epsilon + d\epsilon^2 - f_0(v^-(x,t) - \tau\epsilon m_1) \\ &\quad = -C\epsilon a^+\zeta'(\epsilon(x-Ct)) - f_0(v^-(x,t)) + (\epsilon + \tau\epsilon m_1 + d\epsilon^2). \end{aligned} \tag{2.32}$$

We distinguish among three cases.

Case (i) $\zeta(\epsilon(x-Ct)) < \frac{\delta}{2}$: Clearly, $v^-(x,t) \in \left[-\delta, -\frac{\delta}{2}\right]$ for all $x \in R$ and $t \geq 0$. By (2.32) and (2.27), it follows that

$$\begin{aligned} &\tfrac{\partial v^-}{\partial t} - d_0\Delta v^- - f(v^-(x,t), v^-(x,t-\tau)) \\ &\quad \leq -\min\left\{f_0(u); u \in \left[-\delta, -\tfrac{\delta}{2}\right]\right\} \\ &\qquad +(\epsilon + \tau\epsilon m_1 + d\epsilon^2) < 0. \end{aligned}$$

Case (ii) $\zeta(\epsilon(x-Ct)) > 1-\frac{\delta}{2}$: It then follows that

$$\begin{aligned}1-\delta \geq v^-(x,t) &\geq -\delta + \left(1-\frac{\delta}{2}\right)[1-(1-a^+-2\delta)]\\ &= a^+ + \frac{\delta}{2}[2-(2\delta+a^+)]\\ &\geq a^+ + \frac{\delta}{2}. \qquad (2.33)\end{aligned}$$

Therefore, by (2.32) and (2.28), we have

$$\begin{array}{l}\frac{\partial v^-}{\partial t} - d\Delta v^- - f(v^-(x,t), v^-(x,t-\tau))\\ \leq -\min\left\{f_0(u); u\in\left[a^+ + \frac{\delta}{2}, 1-\delta\right]\right\}\\ \quad +(\epsilon+\tau\epsilon m_1 + d\epsilon^2) < 0.\end{array}$$

Case (iii) $\zeta(\epsilon(x-Ct)) \in \left[\frac{\delta}{2}, 1-\frac{\delta}{2}\right]$: By (2.32) and (2.29), we have

$$\begin{aligned}\frac{\partial v^-}{\partial t} - d_0\Delta v^- - f(v^-(x,t), v^-(x,t-\tau)) \leq{}& -C\epsilon a^+ m_2 + \max\{|f_0(u)|;\\ & u\in[-\delta, 1-\delta]\}\\ & +(\epsilon+\tau\epsilon m_1 + d\epsilon^2) < 0.\end{aligned}$$

Combining cases (i)–(iii), we have

$$\frac{\partial v^-}{\partial t} - d\Delta v^- - f(v^-(x,t), v^-(x,t-\tau)) \leq 0, \quad x\in R, \quad t\geq 0.$$

Then $v^-(x,t)$ is a subsolution of (2.1) on $[0,\infty)$. In a similar way, we can prove $v^+(x,t)$ is a supersolution of (2.1) on $[0,\infty)$.

This completes the proof. □

REMARK 2.2. *It is easy to see $v^+(x,t)$ and $v^-(x,t)$ in Lemma 2.4 have the following properties:*

(P1) *$v^+(x,t) = 1+\delta$ for all $x\geq \xi + C\tau$ and $t\in[-\tau,0]$; $v^+(x,t)\geq a^- - \delta$ for all $x\in R$ and $t\in[-\tau,0]$; $v^+(x,t)\leq \delta + (a-2\delta)e^{-\epsilon t}$ for all $x\leq \xi - Ct - 4\epsilon^{-1}$ and $t\in[0,\infty)$.*

(P2) *$v^-(x,t) = -\delta$ for all $x\leq \xi - C\tau$ and $t\in[-\tau,0]$; $v^-(x,t)\leq a^+ + \delta$ for all $x\in R$ and $t\in[-\tau,0]$; $v^-(x,t)\geq 1-\delta-(1-a^+-2\delta)e^{-\epsilon t}$ for all $x\geq \xi + Ct + 4\epsilon^{-1}$ and $t\in[0,\infty)$.*

LEMMA 2.5. *For any traveling wave solution $U(x-ct)$ of (2.1) with $0\leq U(\xi)\leq 1, \xi\in R$, there holds $\lim_{\xi\to\pm\infty} U'(\xi) = 0$.*

Proof. We prove only $\lim_{\xi\to+\infty} U'(\xi) = 0$ since $\lim_{\xi\to-\infty} U'(\xi) = 0$ follows by the transformation $V(\xi) = U(-\xi)$. By Theorem 2.2, $U(\cdot)\in C^2(R)$ and $U(\xi)$ satisfy (2.20). In the case where $c=0$, (2.20) implies that $U''(\cdot)$ is bounded on R, and hence, $U'(\cdot)$ is uniformly continuous on R. Then, by (2.17) and the Barbalat lemma (see, e.g., [5]), $\lim_{\xi\to+\infty} U'(\xi) = 0$. In the case where $c\neq 0$, we first claim that $\lim_{\xi\to+\infty} U'(\xi)$ exists. Assume that, by contradiction, $\liminf_{\xi\to+\infty} U'(\xi) < \limsup_{\xi\to+\infty} U'(\xi)$. Then, by the fluctuation lemma (see [6]), there are two sequences $\{\xi_n\}_{n=1}^\infty$ and $\{s_n\}_{n=1}^\infty$ with $\xi_n\to+\infty$ and $s_n\to+\infty$ as $n\to+\infty$ such that

$$\lim_{n\to+\infty} U'(\xi_n) = \limsup_{\xi\to+\infty} U'(\xi), \quad U''(\xi_n) = 0, \quad n\geq 1$$

and

$$\lim_{n\to+\infty} U'(s_n) = \liminf_{\xi\to+\infty} U'(\xi), \quad U''(s_n) = 0, \quad n \geq 1.$$

Since $\lim_{\xi\to+\infty} U(\xi) = 1$ and $f(1,1) = 0$, letting $n \to +\infty$ in (2.20) with ξ replaced by ξ_n and s_n, respectively, we get that $\limsup_{\xi\to+\infty} U'(\xi) = 0 = \liminf_{\xi\to+\infty} U'(\xi)$, which is a contradiction to our assumption. Let $\lim_{\xi\to+\infty} U'(\xi) = L$. Since $U(n+1) - U(n) = U'(t_n)$, where $t_n \in [n, n+1]$, letting $n \to \infty$, we have $L = 1 - 1 = 0$.

This completes the proof. □

3. Stability and uniqueness of traveling waves. In this section we will discuss the global asymptotic stability with shift and stability of monotone traveling waves and uniqueness of traveling waves. To prove our main results, we need the following two lemmas.

Let $U(x-ct)$ be a monotone traveling wave solution of (2.1). In view of Lemma 2.3, we define the following two functions:

$$\begin{gathered} w^{\pm}(x,t,\eta,\delta) = U(x - ct + \eta \pm \sigma_0\delta(1 - e^{-\beta_0 t})) \pm \delta e^{-\beta_0 t}, \\ x \in R, \quad t \in [-\tau, \infty), \quad \eta \in R, \text{ and } \delta \in [0, \infty), \end{gathered}$$

where σ_0, β_0 are as in Lemma 2.3. By the proof of Lemma 2.3 we can choose $\beta_0 > 0$ as small as we wish. Then we assume that β_0 has been chosen such that $3e^{\beta_0\tau} < 4$ throughout this section.

LEMMA 3.1. *Let $U(x-ct)$ be a monotone traveling wave solution of* (2.1). *Then there exists a positive number ϵ^* such that, if $u(x,t)$ is a solution of* (2.1) *on* $[0,\infty)$ *with $0 \leq u(x,t) \leq 1$ for $x \in R$ and $t \in [0,\infty)$, and for some $\xi \in R$, $h > 0$, $0 < \delta < \min\left(\bar{\delta}, \frac{1}{\sigma_0}\right)$ and $T \geq 0$, there holds*

$$\begin{gathered} w_0^-(x, -cT + \xi, \delta)(s) \leq u_T(x)(s) \leq w_0^+(x, -cT + \xi + h, \delta)(s), \\ s \in [-\tau, 0], x \in R; \end{gathered}$$

then for every $t \geq T + \tau + 1$, there exist $\hat{\xi}(t), \hat{\delta}(t)$, and $\hat{h}(t)$ such that

$$\begin{gathered} w_0^-(x, -ct + \hat{\xi}(t), \hat{\delta}(t))(s) \leq u_t(x)(s) \leq w_0^+(x, -ct + \hat{\xi}(t) + \hat{h}(t), \hat{\delta}(t))(s), \\ s \in [-\tau, 0], \quad x \in R, \end{gathered}$$

with $\hat{\xi}(t), \hat{\delta}(t)$, and $\hat{h}(t)$ satisfying

$$\hat{\xi}(t) \in [\xi - \sigma_0\delta - 2\sigma_0(\delta + \epsilon^* \min(h,1))e^{\beta_0\tau}, \xi + h + \sigma_0\delta],$$

$$\hat{\delta}(t) = (\delta e^{-\beta_0} + \epsilon^* \min(1,h))e^{-\beta_0(t-(T+\tau+1))},$$

$$\hat{h}(t) \in [0, h + (3e^{\beta_0\tau} - 4)\sigma_0\epsilon^* \min(h,1) + 3e^{\beta_0\tau}\sigma_0\delta].$$

Proof. By Lemma 2.3, $w^+(x,t,-cT+\xi+h,\delta)$ and $w^-(x,t,-cT+\xi,\delta)$ are super- and subsolutions of (2.1), respectively. Clearly, $v(x,t) = u(x,T+t)$, $t \geq 0$, is also a solution of (2.1) with $v_0(x)(s) = u_T(x)(s)$, $s \in [-\tau, 0]$, $x \in R$. Then, by Theorem 2.2, there holds

$$\begin{gathered} w^-(x,t,-cT+\xi,\delta) \leq u(x,T+t) \leq w^+(x,t,-cT+\xi+h,\delta), \\ x \in R, t \geq 0. \end{gathered} \tag{3.1}$$

That is,

$$\begin{aligned} &U(x-c(T+t)+\xi-\sigma_0\delta(1-e^{-\beta_0 t}))-\delta e^{-\beta_0 t}\\ &\quad \leq u(x,T+t)\\ &\quad \leq U(x-c(T+t)+\xi+h+\sigma_0\delta(1-e^{-\beta_0 t}))\\ &\qquad +\delta e^{-\beta_0 t}, \quad x\in R, t\geq 0. \end{aligned} \tag{3.2}$$

Let $z=cT-\xi$. Again by Theorem 2.2, we have that for any $J\geq 0$, all $x\in R$ with $|x-z|\leq J$ and all $t>0$,

$$\begin{aligned} &u(x,T+t)-w^-(x,t,-cT+\xi,\delta)\\ &\quad\geq \theta(J,t)\int_z^{z+1}(u(y,T)-w^-(y,0,-cT+\xi,\delta))dy\\ &\quad=\theta(J,t)\int_z^{z+1}[u(y,T)-(U(y-cT+\xi)-\delta)]dy\\ &\quad=\theta(J,t)\left[\int_z^{z+1}(u(y,T)-U(y-cT+\xi))dy+\delta\right]. \end{aligned} \tag{3.3}$$

By Lemma 2.5, $\lim_{|\eta|\to\infty}U'(\eta)=0$. Then we can fix a positive number M such that $U'(\eta)\leq\frac{1}{2\sigma_0}$ for all $|\eta|\geq M$. Let $J=M+|c|(1+\tau)+1$, $\bar{h}=\min(1,h)$ and

$$\epsilon_1=\frac{1}{2}\min\{U'(x);|x|\leq 2\}>0.$$

Then

$$\int_z^{z+1}[U(y-cT+\xi+\bar{h})-U(y-cT+\xi)]dy=\int_0^1(U(y+\bar{h})-U(y))dy\geq 2\epsilon_1\bar{h},$$

and hence, at least one of the following is true:

(i) $\int_z^{z+1}[u(y,T)-U(y-cT+\xi)]dy\geq\epsilon_1\bar{h}$;

(ii) $\int_z^{z+1}[U(y-cT+\xi+\bar{h})-u(y,T)]dy\geq\epsilon_1\bar{h}$.

In what follows, we consider only the case (i). The case (ii) is similar and thus omitted. For any $s\in[-\tau,0]$, $|x-z|\leq J$, letting $t=1+\tau+s\geq 1$ in (3.3), we have

$$\begin{aligned} u(x,T+1+\tau+s)\geq{}& U(x-z-c(1+\tau+s)-\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)}))\\ &-\delta e^{-\beta_0(1+\tau+s)}+\theta_0(J)\epsilon_1\bar{h}, \end{aligned} \tag{3.4}$$

where $\theta_0(J)=\min_{s\in[-\tau,0]}\theta(J,1+\tau+s)$. Let

$$\begin{aligned} J_1&=J+|c|(1+\tau)+3,\\ \epsilon^*&=\min\left\{\min_{|x|\leq J_1}\frac{\theta_0(J)\epsilon_1}{2\sigma_0U'(x)},\frac{1}{3\sigma_0}\right\}. \end{aligned}$$

By the mean value theorem, it then follows that for all $|x-z|\leq J$, $s\in[-\tau,0]$,

$$\begin{aligned} &U(x-z-c(1+\tau+s)+2\sigma_0\epsilon^*\bar{h}-\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)}))\\ &-U(x-z-c(1+\tau+s)-\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)}))\\ &=U'(\eta_1)2\sigma_0\epsilon^*\bar{h}\leq\theta_0(J)\epsilon_1\bar{h}, \end{aligned} \tag{3.5}$$

and hence,

$$\begin{aligned} u(x,T+1+\tau+s) \geq\ & U(x-c(T+1+\tau+s)+\xi+2\sigma_0\epsilon^*\bar{h} \\ & -\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)}))-\delta e^{-\beta_0(1+\tau+s)}. \end{aligned} \tag{3.6}$$

By the choice of M and J and by the mean value theorem, it then follows that for all $|x-z|\geq J$, $s\in[-\tau,0]$,

$$\begin{aligned} & U(x-c(T+1+\tau+s)+\xi-\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)})) \\ & -U(x-c(T+1+\tau+s)+2\sigma_0\epsilon^*\bar{h}+\xi-\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)})) \\ & = U'(\eta_2)(-2\sigma_0\epsilon^*\bar{h}) \geq -\epsilon^*\bar{h}. \end{aligned} \tag{3.7}$$

That is, for all $|x-z|\geq J$, $s\in[-\tau,0]$,

$$\begin{aligned} & U(x-c(T+1+\tau+s)+\xi-\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)})) \\ & \quad \geq U(x-c(T+1+\tau+s)+\xi+2\sigma_0\epsilon^*\bar{h} \\ & \qquad -\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)}))-\epsilon^*\bar{h}, \end{aligned} \tag{3.8}$$

and hence, by (3.2) with $t=1+\tau+s$, we have

$$\begin{aligned} u(x,T+1+\tau+s) \geq\ & U(x-c(T+1+\tau+s)+\xi+2\sigma_0\epsilon^*\bar{h} \\ & -\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)})) \\ & -\epsilon^*\bar{h}-\delta e^{-\beta_0(1+\tau+s)} \end{aligned} \tag{3.9}$$

for all $|x-z|\geq J,\quad s\in[-\tau,0]$. By (3.6) and (3.9), it follows that for all $x\in R$, $x\in[-\tau,0]$,

$$\begin{aligned} u(x,T+1+\tau+s) \geq\ & U(x-c(T+1+\tau+s)+\xi+2\sigma_0\epsilon^*\bar{h} \\ & -\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)})) \\ & -\delta e^{-\beta_0(1+\tau+s)}-\epsilon^*\bar{h}, \end{aligned} \tag{3.10}$$

and hence,

$$\begin{aligned} u_{T+1+\tau}(x)(s) \geq\ & U(x-cs+[-c(T+1+\tau)+\xi+2\sigma_0\epsilon^*\bar{h} \\ & -\sigma_0\delta(1-e^{-\beta_0(1+\tau+s)})+\sigma_0(\delta e^{-\beta_0}+\epsilon^*\bar{h})(1-e^{-\beta_0 s}) \\ & -\sigma_0(\delta e^{-\beta_0}+\epsilon^*\bar{h})(1-e^{-\beta_0 s})])-(\delta e^{-\beta_0}+\epsilon^*\bar{h})e^{-\beta_0 s} \\ \geq\ & U(x-cs+(-c(T+1+\tau)+\xi+2\sigma_0\epsilon^*\bar{h}+\bar{\xi}) \\ & -\sigma_0(\delta e^{-\beta_0}+\epsilon^*\bar{h})(1-e^{-\beta_0 s})) \\ & -(\delta e^{-\beta_0}+\epsilon^*\bar{h})e^{-\beta_0 s}, \end{aligned} \tag{3.11}$$

where

$$\bar{\xi} = -\sigma_0\delta+\sigma_0(\delta e^{-\beta_0}+\epsilon^*\bar{h})(1-e^{\beta_0\tau}). \tag{3.12}$$

Then

$$u_{T+1+\tau}(x)(s) \geq w_0^-(x,\eta,\delta e^{-\beta_0}+\epsilon^*\bar{h})(s), \quad x\in R, s\in[-\tau,0], \tag{3.13}$$

where $\eta=-c(T+1+\tau)+\xi+2\sigma_0\epsilon^*\bar{h}+\bar{\xi}$. Again, by Theorem 2.2,

$$u_{T+1+\tau+t'}(x)(s) \geq w_{t'}^-(x,\eta,\delta e^{-\beta_0}+\epsilon^*\bar{h})(s) \quad \text{for all } t'\geq 0, s\in[-\tau,0]. \tag{3.14}$$

Then for any $t \geq T + 1 + \tau$, setting $t' = t - (T + 1 + \tau)$ in (3.14), we have

$$\begin{aligned} u_t(x)(s) &\geq w^-_{t-(T+1+\tau)}(x, \eta, \delta e^{-\beta_0} + \epsilon^* \bar{h})(s) \\ &= U(x - cs - ct + c(T + 1 + \tau) + \eta - \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) \cdot \\ &\quad (1 - e^{-\beta_0(t-(T+1+\tau))} \cdot e^{-\beta_0 s})) \\ &\quad -(\delta e^{-\beta_0} + \epsilon^* \bar{h}) e^{-\beta_0(t-(T+1+\tau))} \cdot e^{-\beta_0 s} \\ &\geq U(x - cs - ct + c(T + 1 + \tau) + \eta - \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) \cdot 1 \\ &\quad +\sigma_0 \hat{\delta}(t)(1 - e^{-\beta_0 s}) - \sigma_0 \hat{\delta}(t)(1 - e^{-\beta_0 s})) - \hat{\delta}(t) e^{-\beta_0 s} \\ &\geq U(x - cs - ct + c(T + 1 + \tau) + \eta - \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) \\ &\quad +\sigma_0 \hat{\delta}(t)(1 - e^{\beta_0 \tau})] - \sigma_0 \hat{\delta}(t)(1 - e^{-\beta_0 s})) - \hat{\delta}(t) \cdot e^{-\beta_0 s}, \end{aligned} \tag{3.15}$$

where $\hat{\delta}(t) = (\delta e^{-\beta_0} + \epsilon^* \bar{h}) \cdot e^{-\beta_0(t-(T+1+\tau))}$. By the monotonicity of $U(\cdot)$, the choice of η and (3.12), it then follows that

$$u_t(x)(s) \geq w^-_0(x, -ct + \hat{\xi}(t), \hat{\delta}(t))(s), \quad s \in [-\tau, 0], x \in R, \tag{3.16}$$

where

$$\begin{aligned} \hat{\xi}(t) &= \xi + 2\sigma_0 \epsilon^* \bar{h} + [-\sigma_0 \delta + \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h})(1 - e^{\beta_0 \tau})] \\ &\quad -\sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) + \sigma_0 \hat{\delta}(t)(1 - e^{\beta_0 \tau}) \\ &= \xi + 2\sigma_0 \epsilon^* \bar{h} - \sigma_0 \delta - \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) e^{\beta_0 \tau} \\ &\quad +\sigma_0 \hat{\delta}(t) - \sigma_0 \hat{\delta}(t) e^{\beta_0 \tau}. \end{aligned}$$

Therefore it follows that

$$\begin{aligned} \hat{\xi}(t) &\geq \xi - \sigma_0 \delta - \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) e^{\beta_0 \tau} - \sigma_0 \hat{\delta}(t) e^{\beta_0 \tau} \\ &\geq \xi - \sigma_0 \delta - 2\sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) e^{\beta_0 \tau} \\ &\geq \xi - \sigma_0 \delta - 2\sigma_0(\delta + \epsilon^* \bar{h}) e^{\beta_0 \tau}, \end{aligned} \tag{3.17}$$

and, by the choice of ϵ^*,

$$\begin{aligned} \hat{\xi}(t) &\leq \xi + 2\sigma_0 \epsilon^* \bar{h} + \sigma_0(\delta e^{-\beta_0} + \epsilon^* \bar{h}) \\ &\leq \xi + 3\sigma_0 \epsilon^* \bar{h} + \sigma_0 \delta \\ &\leq \xi + h + \sigma_0 \delta. \end{aligned} \tag{3.18}$$

For any $t \geq T$, by the inequality of the right-hand side of (3.2), we have

$$u(x, t) \leq U(x - ct + \xi + h + \sigma_0 \delta(1 - e^{-\beta_0(t-T)})) + \delta e^{-\beta_0(t-T)}. \tag{3.19}$$

It then follows that, for all $t \geq T + 1 + \tau$,

$$\begin{aligned} u_t(x)(s) = u(x, t + s) \leq U(x &- c(t + s) + \xi + h \\ &+\sigma_0 \delta(1 - e^{-\beta_0(t+s-T)})) + \delta e^{-\beta_0(t+s-T)}, \\ &s \in [-\tau, 0], x \in R. \end{aligned} \tag{3.20}$$

Therefore for all $t \geq T + 1 + \tau$,

$$\begin{aligned} u_t(x)(s) &\leq U(x - cs - ct + \xi + h + \sigma_0 \delta(1 - e^{-\beta_0(t+s-T)})) + \hat{\delta}(t) e^{-\beta_0 s} \\ &= U(x - cs - ct + \xi + h + \sigma_0 \delta(1 - e^{-\beta_0(t-T)} \cdot e^{-\beta_0 s}) \\ &\quad -\sigma_0 \hat{\delta}(t)(1 - e^{-\beta_0 s}) + \sigma_0 \hat{\delta}(t)(1 - e^{-\beta_0 s})) + \hat{\delta}(t) e^{-\beta_0 s} \\ &\leq U(x - cs - ct + \xi + h + \sigma_0 \delta - \sigma_0 \hat{\delta}(t)(1 - e^{\beta_0 \tau}) \\ &\quad +\sigma_0 \hat{\delta}(t)(1 - e^{-\beta_0 s})) + \hat{\delta}(t) e^{-\beta_0 s}, \quad s \in [-\tau, 0], x \in R. \end{aligned} \tag{3.21}$$

It then follows that for all $t \geq T + 1 + \tau$,

$$u_t(x)(s) \leq w_0^+(x, -ct + \hat{\xi}(t) + \hat{h}(t), \hat{\delta}(t))(s), \quad s \in [-\tau, 0], x \in R, \tag{3.22}$$

where

$$\begin{aligned}\hat{h}(t) &= \xi + h + \sigma_0\delta - \sigma_0\hat{\delta}(t)(1 - e^{\beta_0\tau}) - \hat{\xi}(t)\\ &= h - 2\sigma_0\epsilon^*\bar{h} + 2\sigma_0\delta + 2\sigma_0\hat{\delta}(t)(e^{\beta_0\tau} - 1)\\ &\quad +\sigma_0(\delta e^{-\beta_0} + \epsilon^*\bar{h})e^{\beta_0\tau}.\end{aligned}\tag{3.23}$$

Notice that we have used the expression of $\hat{\xi}(t)$ in getting the second equality of (3.23). By the choice of ϵ^*, we have $h - 2\sigma_0\epsilon^*\bar{h} \geq h - 2\sigma_0\epsilon^* h = (1 - 2\sigma_0\epsilon^*)h > 0$, and hence,

$$\begin{aligned}0 < \hat{h}(t) &\leq h - 2\sigma_0\epsilon^*\bar{h} + 2\sigma_0\delta + 2\sigma_0(\delta e^{-\beta_0} + \epsilon^*\bar{h}) \cdot (e^{\beta_0\tau} - 1)\\ &\quad +\sigma_0(\delta e^{-\beta_0} + \epsilon^*\bar{h})e^{\beta_0\tau}\\ &\leq h - 2\sigma_0\epsilon^*\bar{h} + 2\sigma_0\delta + \sigma_0(\delta e^{-\beta_0} + \epsilon^*\bar{h}) \cdot (3e^{\beta_0\tau} - 2)\\ &\leq h - 2\sigma_0\epsilon^*\bar{h} + 2\sigma_0\delta + \sigma_0(\delta + \epsilon^*\bar{h})(3e^{\beta_0\tau} - 2)\\ &= h + (3e^{\beta_0\tau} - 4)\sigma_0\epsilon^*\bar{h} + 3e^{\beta_0\tau} \cdot \sigma_0\delta.\end{aligned}\tag{3.24}$$

Combining (3.16) and (3.22), now we complete the proof. □

LEMMA 3.2. *Let $U(x - ct)$ be a monotone traveling wave solution of* (2.1)*, and let $\varphi \in [0, 1]_C$ be such that*

$$\liminf_{x\to\infty} \min_{s\in[-\tau,0]} \varphi(x, s) > a^+, \quad \limsup_{x\to-\infty} \max_{s\in[-\tau,0]} \varphi(x, s) < a^-.$$

Then, for any $\delta > 0$, there exist $T = T(\varphi, \delta) > 0$, $\xi = \xi(\varphi, \delta) \in R$, and $h = h(\varphi, \delta) > 0$ such that

$$\begin{gathered}w_0^-(x, -cT + \xi, \delta)(s) \leq u_T(x, \varphi)(s) \leq w_0^+(x, -cT + \xi + h, \delta)(s),\\ s \in [-\tau, 0], x \in R.\end{gathered}$$

Proof. By Theorem 2.2, $u(x, t, \varphi)$ exists globally on $[0, \infty)$ and $0 \leq u(x, t, \varphi) \leq 1$, $x \in R$, $t \geq 0$. For any $\delta > 0$, we choose a $0 < \delta_1 = \delta_1(\delta, \varphi) < \min(\delta, \bar{\delta}_0)$ such that

$$\liminf_{x\to\infty} \min_{s\in[-\tau,0]} \varphi(x, s) > a^+ + \delta_1,$$

$$\limsup_{x\to-\infty} \max_{s\in[-\tau,0]} \varphi(x, s) < a^- - \delta_1.$$

Then there exists $M = M(\varphi, \delta_1) > 0$ such that

$$\begin{aligned}&\varphi(x, s) \leq a^- - \delta_1 \quad \text{for all } x \leq -M \text{ and } s \in [-\tau, 0],\\ &\varphi(x, s) \geq a^+ + \delta_1 \quad \text{for all } x \geq M \text{ and } s \in [-\tau, 0].\end{aligned}\tag{3.25}$$

Let $\epsilon = \epsilon(\delta_1)$ and $C = C(\delta_1)$ be defined in Lemma 2.4 with δ replaced by δ_1. Define $\xi^+ = -M - C\tau$ and $\xi^- = M + C\tau$, and let $v^\pm(x, t)$ be defined in Lemma 2.4 with $\xi = \xi^\pm$, respectively. By (3.25) and Remark 2.2, it follows that for all $s \in [-\tau, 0]$,

$$\begin{aligned}&\varphi(x, s) \leq a^- - \delta_1 \leq v^+(x, s) &&\text{for } x \leq -M,\\ &\varphi(x, s) \leq 1 < 1 + \delta_1 = v^+(x, s) &&\text{for } x \geq \xi^+ + C\tau = -M\end{aligned}$$

and

$$\begin{aligned}&\varphi(x,s) \geq a^{+} + \delta_1 \geq v^{-}(x,s) \quad \text{for } x \geq M,\\ &\varphi(x,s) \geq 0 > -\delta_1 = v^{-}(x,s) \quad \text{for } x \leq \xi^{-} - C\tau = M.\end{aligned}$$

Then we have

$$v^{-}(x,s) \leq \varphi(x,s) \leq v^{+}(x,s), \quad x \in R, s \in [-\tau, 0]. \tag{3.26}$$

By Lemma 2.4 and Theorem 2.2, it follows that

$$v^{-}(x,t) \leq u(x,t,\varphi) \leq v^{+}(x,t), \quad x \in R, t \geq 0. \tag{3.27}$$

Since $\delta_1 < \delta$, we choose a sufficiently large positive number $T > \tau$ such that, for all $t \geq T - \tau$,

$$\delta_1 + (a^{-} - 2\delta_1)e^{-\epsilon t} < \delta \quad \text{and } 1 - \delta_1 - (1 - a^{+} - 2\delta_1)e^{-\epsilon t} > 1 - \delta,$$

and hence, again by Remark 2.2,

$$u(x,t,\varphi) \leq v^{+}(x,t) < \delta \quad \text{for } x \leq \xi^{+} - Ct - 4\epsilon^{-1} \tag{3.28}$$

and

$$u(x,t,\varphi) \geq v^{-}(x,t) > 1 - \delta \quad \text{for } x \geq \xi^{-} + Ct + 4\epsilon^{-1}. \tag{3.29}$$

Let $x^{-} = \xi^{+} - CT - 4\epsilon^{-1}$ and $x^{+} = \xi^{-} + CT + 4\epsilon^{-1}$. By (3.28) and (3.29), it follows that for any $t \in [T - \tau, T]$,

$$\begin{aligned}&u(x,t,\varphi) < \delta \quad \text{for } x \leq x^{-}\\ &u(x,t,\varphi) > 1 - \delta \quad \text{for } x \geq x^{+}.\end{aligned} \tag{3.30}$$

By (2.17), there exists a sufficiently large positive number H such that $\frac{H}{2} > x^{+}$, $-\frac{H}{2} < x^{-}$, and

$$U(x) + \delta > 1 \quad \text{for all } x \geq \frac{H}{2} \quad \text{and } U(x) - \delta < 0 \quad \text{for all } x \leq -\frac{H}{2}. \tag{3.31}$$

Combining (3.30), (3.31), and the fact that $0 \leq U(x) \leq 1$ and $0 \leq u(x,t,\varphi) \leq 1$ for all $x \in R$ and $t \in [0, \infty)$, we then have

$$U(-H+x) - \delta \leq u(x,t,\varphi) \leq U(H+x) + \delta \quad \text{for all } x \in R \quad \text{and } t \in [T-\tau, T]. \tag{3.32}$$

Let $\xi_0 = -H + cT$, $h_0 = 2H > 0$. Then (3.32) implies that

$$U(x - cT + \xi_0) - \delta \leq u_T(x,\varphi)(s) \leq U(x - cT + \xi_0 + h_0) + \delta, \quad s \in [-\tau, 0], \quad x \in R. \tag{3.33}$$

Let $\xi = \xi_0 + \sigma_0\delta(1 - e^{\beta_0\tau}) - |c|\tau$ and $h = h_0 - 2\sigma_0\delta(1 - e^{\beta_0\tau}) + 2|c|\tau > 0$. Then by (3.33), we have that for any $s \in [-\tau, 0]$ and $x \in R$,

$$\begin{aligned}&w_0^{-}(x, -cT + \xi, \delta)(s)\\ &\quad = U(x - cs - cT + \xi - \sigma_0\delta(1 - e^{-\beta_0 s})) - \delta e^{-\beta_0 s}\\ &\quad \leq U(x + |c|\tau - cT + \xi - \sigma_0\delta(1 - e^{\beta_0\tau})) - \delta\\ &\quad = U(x - cT + \xi_0) - \delta \leq u_T(x,\varphi)(x)\end{aligned}$$

and

$$\begin{aligned}&w_0^+(x, -cT + \xi + h, \delta)(s)\\ &\quad = U(x - cs - cT + \xi + h + \sigma_0\delta(1 - e^{-\beta_0 s})) + \delta e^{-\beta_0 s}\\ &\quad \geq U(x - |c|\tau - cT + \xi + h + \sigma_0\delta(1 - e^{\beta_0\tau})) + \delta\\ &\quad = U(x - cT + \xi_0 + h_0) + \delta \geq u_T(x, \varphi)(s).\end{aligned}$$

It then follows that

$$w_0^-(x, -cT + \xi, \delta)(s) \leq u_T(x, \varphi)(s) \leq w_0^+(x, -cT + \xi + h, \delta)(s).$$

This completes the proof. □

Now we are in a position to prove the main results in this section.

THEOREM 3.3. *Assume that* (2.1) *has a monotone traveling wave solution* $U(x - ct)$. *Then* $U(x - ct)$ *is globally asymptotically stable with phase shift in the sense that there exists* $k > 0$ *such that for any* $\varphi \in [0, 1]_C$ *with*

$$\liminf_{x\to\infty} \min_{s\in[-\tau,0]} \varphi(x, s) > a^+, \quad \limsup_{x\to-\infty} \max_{s\in[-\tau,0]} \varphi(x, s) < a^-;$$

the solution $u(x, t, \varphi)$ *of* (2.1) *satisfies*

$$|u(x, t, \varphi) - U(x - ct + \xi)| \leq Ke^{-kt}, \quad x \in R, t \geq 0$$

for some $K = K(\varphi) > 0$ *and* $\xi = \xi(\varphi) \in R$.

Proof. Let β_0, σ_0, $\bar{\delta}$ be as in Lemma 2.3 with β_0 chosen such that $3e^{\beta_0\tau} < 4$, and then let ϵ^* be as in Lemma 3.1 with ϵ^* chosen such that $(4 - 3e^{\beta_0\tau})\sigma_0 \cdot \epsilon^* < 1$. We further choose a $0 < \delta^* < \min(\frac{\delta_0}{2}, \bar{\delta}, \frac{1}{\sigma_0})$ such that

$$1 > k^* := (4 - 3e^{\beta_0\tau})\sigma_0\epsilon^* - 3e^{\beta_0\tau}\sigma_0\delta^* > 0,$$

and then fix a $t^* \geq \tau + 1$ such that

$$e^{-\beta_0(t^*-\tau-1)}(1 + \epsilon^*/\delta^*) < 1 - k^*.$$

We first prove the following two claims.

Claim 1. There exist $T^* = T^*(\varphi) > 0$, $\xi^* = \xi^*(\varphi) \in R$ such that

$$\begin{aligned}&w_0^-(x, -cT^* + \xi^*, \delta^*)(s) \leq u_{T^*}(x, \varphi)(s) \leq w_0^+(x, -cT^* + \xi^* + 1, \delta^*)(s),\\ &\qquad\qquad s \in [-\tau, 0], x \in R.\end{aligned} \tag{3.34}$$

Indeed, by Lemma 3.2, there exist $T = T(\varphi) > 0$, $\xi = \xi(\varphi) \in R$, and $h = h(\varphi) > 0$ such that

$$w_0^-(x, -cT + \xi, \delta^*)(s) \leq u_T(x, \varphi)(s) \leq w_0^+(x, -cT + \xi + h, \delta^*)(s), \tag{3.35}$$

$$s \in [-\tau, 0], \quad x \in R.$$

If $h \leq 1$, (3.34) follows immediately from the monotonicity of $U(\cdot)$. Then we assume that $h > 1$ and let

$$N = \max\{m; m \text{ is a nonnegative integer and } mk^* < h\}.$$

Since $0 < k^* < 1$ and $h > 1$, we have $N \geq 1$, $Nk^* < h \leq (N+1)k^*$, and hence, $0 < h - Nk^* \leq (N+1)k^* - Nk^* = k^* < 1$. Clearly, $\bar{h} := \min(1, h) = 1$. By (3.35), Lemma 3.1 and the choice of t^* and k^*, it then follows that

$$\begin{aligned} w_0^-(x, -c(T+t^*) + \hat{\xi}(T+t^*), \hat{\delta}(T+t^*))(s) &\leq u_{T+t^*}(x, \varphi)(s) \\ &\leq w_0^+(x, -c(T+t^*) + \hat{\xi}(T+t^*) \\ &\quad + \hat{h}(T+t^*), \hat{\delta}(T+t^*))(s), \\ &\quad\ s \in [-\tau, 0], x \in R, \end{aligned} \tag{3.36}$$

where

$$\begin{aligned} &\hat{\xi}(T+t^*) \in [\xi - \sigma_0\delta^* - 2\sigma_0(\delta^* + \epsilon^*)e^{\beta_0\tau}, \xi + h + \sigma_0\delta^*], \\ &\hat{\delta}(T+t^*) = (\delta^* e^{-\beta_0} + \epsilon^*)e^{-\beta_0(t^*-\tau-1)} \leq (1-k^*)\delta^* < \delta^*, \\ &0 \leq \hat{h}(T+t^*) \leq h + (3e^{\beta_0\tau} - 4)\sigma_0\epsilon^* + 3e^{\beta_0\tau}\sigma_0\delta^* = h - k^*. \end{aligned}$$

Repeating the same process N times, we have that (3.36), with $T + t^*$ replaced by $T + Nt^*$, holds for some $\hat{\xi} \in R$, $0 < \hat{\delta} \leq \delta^*$, and $0 \leq \hat{h} \leq h - Nk^* < 1$. Let $T^* = T + Nt^*$, $\xi^* = \hat{\xi}$. Again, by the monotonicity of $U(\cdot)$, (3.34) then follows.

Claim 2. Let $p = 2\sigma_0\delta^* + 2\sigma_0(\delta^* + \epsilon^*)e^{\beta_0\tau} + 1$, $T_m = T^* + mt^*$, $\delta_m^* = (1-k^*)^m\delta^*$, and $h_m = (1-k^*)^m$, $m \geq 0$. Then there exists a sequence $\{\xi_m\}_{m=0}^{\infty} \subset R$ with $\xi_0 = \xi^*$ such that

$$|\xi_{m+1} - \xi_m| \leq ph_m, \quad m \geq 0 \tag{3.37}$$

and

$$\begin{aligned} w_0^-(x, -cT_m + \xi_m, \delta_m^*)(s) &\leq u_{T_m}(x, \varphi)(s) \\ &\leq w_0^+(x, -cT_m + \xi_m + h_m, \delta_m^*)(s), \\ &\quad\ s \in [-\tau, 0], \quad x \in R, m \geq 0. \end{aligned} \tag{3.38}$$

In fact, Claim 1 implies that (3.38) holds for $m = 0$. Now suppose that (3.38) holds for some $m = l \geq 0$. By Lemma 3.1 with $T = T_l$, $\xi = \xi_l$, $h = h_l$, $\delta = \delta_l^*$, and $t = T_l + t^* = T_{l+1}$ (since $t^* \geq \tau + 1$), we then have

$$w_0^-(x, -cT_{l+1} + \hat{\xi}, \hat{\delta})(s) \leq u_{T_{l+1}}(x, \varphi)(s) \leq w_0^+(x, -cT_{l+1} + \hat{\xi} + \hat{h}, \hat{\delta})(s), \tag{3.39}$$

$$s \in [-\tau, 0], \quad x \in R,$$

where

$$\begin{aligned} \hat{\xi} &\in [\xi_l - \sigma_0\delta_l^* - 2\sigma_0(\delta_l^* + \epsilon^* h_l)e^{\beta_0\tau},\ \xi_l + h_l + \sigma_0\delta_l^*], \\ \hat{\delta} &= (\delta_l e^{-\beta_0} + \epsilon^* h_l)e^{-\beta_0(T_{l+1} - T_l - \tau - 1)} \\ &\quad \leq (1-k^*)^l \cdot \delta^* \left[\left(1 + \frac{\epsilon^*}{\delta^*}\right) e^{-\beta_0(t^*-\tau-1)}\right] \\ &\quad \leq (1-k^*)^l \cdot \delta^*(1-k^*) = \delta_{l+1}^*, \\ \hat{h} &\leq h_l - (4 - 3e^{\beta_0\tau})\sigma_0\epsilon^* h_l + 3e^{\beta_0\cdot\tau}\sigma_0\ \delta_l \\ &\quad = (1-k^*)^l \cdot [1 - (4 - 3e^{\beta_0\tau})\sigma_0\epsilon^* + 3e^{\beta_0\tau}\sigma_0\delta^*] \\ &\quad = (1-k^*)^{l+1} = h_{l+1}. \end{aligned}$$

We choose $\xi_{l+1} = \hat{\xi}$. Then

$$|\xi_{l+1} - \xi_l| \leq |\xi_l + h_l + \sigma_0\delta_l^* - (\xi_l - \sigma_0\delta_l^* - 2\sigma_0(\delta_l^* + \epsilon^* h_l) \cdot e^{\beta_0\tau})| = p \cdot h_l.$$

It then follows that (3.37) holds for $m = l$ and (3.38) holds for $m = l+1$. By induction, (3.37) and (3.38) hold for all $m \geq 0$.

For each $m \geq 0$, by (3.38) and Theorem 2.2, it follows that for all $t \geq T_m$, $x \in R$,

$$\begin{aligned} &U(x - ct + \xi_m - \sigma_0\delta_m^*(1 - e^{-\beta_0(t-T_m)})) - \delta_m^* e^{-\beta_0(t-T_m)} \leq u(x, t, \varphi) \\ &\leq U(x - ct + \xi_m + h_m + \sigma_0\delta_m^*(1 - e^{-\beta_0(t-T_m)})) + \delta_m^* e^{-\beta_0(t-T_m)}. \end{aligned} \tag{3.40}$$

For any $t \geq T^*$, let $m = [\frac{t-T^*}{t^*}] \geq 0$ be the largest integer not greater than $\frac{t-T^*}{t^*}$, and define $\delta(t) = \delta_m^*$, $\xi(t) = \xi_m - \sigma_0\delta_m^*$, and $h(t) = h_m + 2\sigma_0\delta_m^*$; then we have $T_m = T^* + mt^* \leq t < T^* + (m+1)t^* = T_{m+1}$. By (3.40), it follows that for all $t \geq T^*$, $x \in R$,

$$U(x - ct + \xi(t)) - \delta(t) \leq u(x, t, \varphi) \leq U(x - ct + \xi(t) + h(t)) + \delta(t). \tag{3.41}$$

Moreover, we have

$$\delta(t) = \delta_m^* = [1 - k^*]^m \delta^* \leq \delta^* \exp\left\{\left(\frac{t - T^*}{t^*} - 1\right)\ln(1 - k^*)\right\}, \quad t \geq T^*, \tag{3.42}$$

$$\begin{aligned} h(t) = h_m + 2\sigma_0\delta_m^* &= (1 + 2\sigma_0\delta^*)(1 - k^*)^m \\ &\leq (1 + 2\sigma_0\delta^*)\exp\left\{\left(\frac{t - T^*}{t^*} - 1\right)\ln(1 - k^*)\right\}, \quad t \geq T^*, \end{aligned} \tag{3.43}$$

and for any $r \geq t \geq T^*$, by (3.37),

$$\begin{aligned} |\xi(r) - \xi(t)| &= |\xi_m - \sigma_0\delta_m^* - (\xi_n - \sigma_0\delta_n^*)| \\ &\leq |\xi_m - \xi_n| + \sigma_0\delta_m^* + \sigma_0\delta_n^* \\ &\leq \sum_{l=n}^{m-1} |\xi_{l+1} - \xi_l| + 2\sigma_0\delta_n^* \\ &\leq \sum_{l=n}^{m-1} p \cdot h_l + 2\sigma_0\delta_n^* \\ &\leq \frac{ph_n}{1 - (1 - k^*)} + 2\sigma_0\delta_n^* \\ &= q \cdot \delta(t), \end{aligned} \tag{3.44}$$

where $m = \left[\frac{r-T^*}{t^*}\right] \geq n = \left[\frac{t-T^*}{t^*}\right]$ and $q = \frac{p}{k^*\delta^*} + 2\sigma_0$. Clearly, (3.44) implies that $\xi(\infty) = \lim_{t\to\infty} \xi(t)$ exists and

$$|\xi(\infty) - \xi(t)| \leq q\delta(t), \quad t \geq T^*.$$

Then

$$|\xi(t) - \xi(\infty)| \leq q \cdot \delta^* \exp\left\{\left(\frac{t - T^*}{t^*} - 1\right)\ln(1 - k^*)\right\}, \quad t \geq T^*. \tag{3.45}$$

Therefore, by defining $k = -\frac{1}{t^*}\ln(1-k^*) > 0$ and combining (3.41), (3.42), (3.43), and (3.45), we obtain the assertion of the theorem.

This completes the proof. □

THEOREM 3.4. *Every monotone traveling wave solution of* (2.1) *is Liapunov stable. If* (2.1) *has a monotone traveling wave solution $U(x-ct)$, then the traveling wave solutions of* (2.1) *are unique up to a translation in the sense that for any traveling wave solution $\bar{U}(x-\bar{c}t)$ with $0 \le \bar{U}(\xi) \le 1$, $\xi \in R$, we have $\bar{c} = c$ and $\bar{U}(\cdot) = U(\xi_0 + \cdot)$ for some $\xi_0 = \xi_0(\bar{U}) \in R$.*

Proof. Let $U(x-ct)$ be a monotone traveling wave solution of (2.1). By the uniform continuity of $U(\cdot)$ on R, it follows that for any $\epsilon > 0$, there exists a $\delta_1 = \delta_1(\epsilon) > 0$ such that for all $|y| \le \delta_1$,

$$|U(x-ct+y) - U(x-ct)| < \frac{\epsilon}{2}, \quad x \in R, t \ge 0. \tag{3.46}$$

We then further choose a $\delta = \delta(\epsilon) > 0$ such that $\delta < \min(\frac{\epsilon}{2}, \frac{\delta_1 \cdot e^{-\beta_0\tau}}{\sigma_0}, \bar{\delta})$, where β_0, σ_0, and $\bar{\delta}$ are as in Lemma 2.3. For any $\varphi \in C([-\tau,0],X)$ with $|\varphi(x,s) - U(x-cs)| < \delta$ for $s \in [-\tau,0]$ and $x \in R$, we have

$$\begin{aligned} &U(x-cs+\sigma_0\delta(1-e^{\beta_0\tau}) - \sigma_0\delta(1-e^{-\beta_0 s})) - \delta e^{-\beta_0 s} \\ &\quad \le \varphi(x,s) \le U(x-cs+\sigma_0\delta(e^{\beta_0\tau}-1) \\ &\qquad +\sigma_0\delta(1-e^{-\beta_0 s})) + \delta e^{-\beta_0 s}, \quad s \in [-\tau,0], x \in R. \end{aligned} \tag{3.47}$$

By Lemma 2.3 and Theorem 2.2, it then follows that

$$\begin{aligned} &U(x-ct+\sigma_0\delta(1-e^{\beta_0\tau}) - \sigma_0\delta(1-e^{-\beta_0 t})) - \delta e^{-\beta_0 t} \le u(x,t,\varphi) \\ &\quad \le U(x-ct+\sigma_0\delta(e^{\beta_0\tau}-1) + \sigma_0\delta(1-e^{-\beta_0 t})) \\ &\qquad +\delta e^{-\beta_0 t}, x \in R, t \ge 0. \end{aligned} \tag{3.48}$$

By the choice of $\delta = \delta(\epsilon)$, we have that for all $t \ge 0$,

$$\begin{aligned} |\sigma_0\delta(1-e^{\beta_0\tau}) - \sigma_0\delta(1-e^{-\beta_0 t})| &\le \sigma_0\delta(e^{\beta_0\tau}-1) + \sigma_0\delta(1-e^{-\beta_0 t}) \\ &\le \sigma_0\delta e^{\beta_0\tau} < \delta_1(\epsilon) \end{aligned}$$

and

$$|\sigma_0\delta(e^{\beta_0\tau}-1) + \sigma_0\delta(1-e^{-\beta_0 t})| \le \sigma_0\delta e^{\beta_0\tau} < \delta_1(\epsilon).$$

Then, by (3.46) and (3.48), it follows that $U(x-ct) - \epsilon \le u(x,t,\varphi) \le U(x-ct) + \epsilon$, $x \in R$, $t \ge 0$. That is, $|u(x,t,\varphi) - U(x-ct)| < \epsilon$, $x \in R$, $t \ge 0$. Therefore $U(x-ct)$ is Liapunov stable.

To prove the uniqueness, we let $U(x-ct)$ be the given monotone traveling wave solution, and let $\bar{U}(x-\bar{c}t)$ be any traveling wave solution of (2.1) with $0 \le \bar{U} \le 1$ on R. Since $\lim_{x\to\infty}\bar{U}(x-\bar{c}s) = 1$ and $\lim_{x\to-\infty}\bar{U}(x-\bar{c}s) = 0$ uniformly for $s \in [-\tau,0]$, we have

$$\begin{aligned} &\liminf_{x\to\infty} \min_{s\in[-\tau,0]} \bar{U}(x-\bar{c}s) > a^+ \quad \text{and} \\ &\limsup_{x\to-\infty} \max_{s\in[-\tau,0]} \bar{U}(x-\bar{c}s) < a^-. \end{aligned} \tag{3.49}$$

Then, by Theorem 3.3, there exists $K_0 = K_0(\bar{U}) > 0$ and $\xi_0 = \xi_0(\bar{U}) \in R$ such that

$$|\bar{U}(x-\bar{c}t) - U(x-ct+\xi_0)| \le K_0 e^{-kt}, \quad x \in R, \quad t \ge 0. \tag{3.50}$$

Let $\bar{\xi} \in R$ be such that $0 < \bar{U}(\bar{\xi}) < 1$, and define $L(\bar{\xi}) := \{(x,t); x \in R, t \geq 0, x - \bar{c}t = \bar{\xi}\}$. By (3.50), it then follows that

$$U(\bar{\xi} + \xi_0 + (\bar{c} - c)t) - K_0 e^{-kt} \leq \bar{U}(\bar{\xi}) \leq U(\bar{\xi} + \xi_0 + (\bar{c} - c)t) + K_0 e^{-kt} \quad \text{for all } (x,t) \in L(\bar{\xi}). \tag{3.51}$$

Since $U(\infty) = 1$ and $U(-\infty) = 0$, letting $t \to \infty$ in (3.51), we obtain that $\bar{c} \leq c$ from the first inequality and that $\bar{c} \geq c$ from the second inequality. Then $\bar{c} = c$. For any $\xi \in R$, again by (3.50), we then have

$$|\bar{U}(\xi) - U(\xi + \xi_0)| \leq K_0 e^{-kt} \quad \text{for all } (x,t) \in L(\xi). \tag{3.52}$$

Therefore, letting $t \to \infty$ in (3.52), we get $\bar{U}(\xi) = U(\xi + \xi_0)$ for all $\xi \in R$, that is, $\bar{U}(\cdot) = U(\xi_0 + \cdot)$.

This completes the proof. □

Remark 3.1. *For the typical Huxley nonlinearity*

$$f(u,v) = \begin{cases} u(1-u)(v-a) & \textit{for } 0 \leq u \leq 1, v \in R, \\ u(1-u)(u-a) & \textit{otherwise}, \end{cases}$$

with $0 < a < 1$, *let* $\hat{f} : I^2 \to R$ *be a smooth extension of* $f : [0,1]^2 \to R$ *such that* (H1) *and* (H2) *hold for* $\hat{f}$. *Clearly,* [7, Corollary 5] *implies that* $[0,1]_C$ *is positively invariant for* (2.1) *with* f *replaced by* $\hat{f}$. *It then follows that Theorems* 3.3 *and* 3.4 *hold for* (2.1) *with the Huxley nonlinearity.*

Remark 3.2. *With some additional assumptions on* $f(\cdot,\cdot)$, *Schaaf* [13] *proved the existence of the monotone traveling wave solution of* (2.1) *and uniqueness of the wave speeds. By Theorems* 3.3 *and* 3.4 *we further conclude that this monotone traveling wave solution is globally and exponentially asymptotically stable with phase shift, that all traveling waves are unique up to translation, and that every traveling wave solution is Liapunov stable.*

Acknowledgment. The authors would like to thank Professor H. Matano for valuable discussions during his visit to the Arizona State University.

REFERENCES

[1] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[2] D. Daners and P. K. Medina, *Abstract Evolution Equations, Periodic Problems and Applications*, Pitman Res. Notes Math. Ser. 279, Longman Scientific & Technical, Harlow, 1992.

[3] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to traveling front solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.

[4] H. I. Freedman and X.-Q. Zhao, *Global asymptotics in some quasimonotone reaction-diffusion systems with delays*, J. Differential Equations, 137 (1997), pp. 340–362.

[5] K. Gopalsamy, *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, Kluwer Academic Publishers, Dordrecht, 1992.

[6] M. Hirsch, H. Hanish, and J.-P. Gabriel, *Differential equation model of some parasitic infections: Methods for the study of asymptotic behavior*, Comm. Pure. Appl. Math., 38 (1985), pp. 733–753.

[7] R. H. Martin and H. L. Smith, *Abstract functional-differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp. 1–44.

[8] R. H. Martin and H. L. Smith, *Reaction-diffusion systems with time delays: Monotonicity, invariance, comparison and convergence*, J. Reine Angew. Math., 413 (1991), pp. 1–35.

[9] T. Ogiwara and H. Matano, *Monotonicity and convergence results in order-preserving systems in the presence of symmetry*, Discrete Contin. Dynam. Systems, 5 (1999), pp. 1–34.

[10] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[11] J.-M. Roquejoffre, D. Terman, and V. A. Volpert, *Global stability of traveling fronts and convergence towards stacked families of waves in monotone parabolic systems*, SIAM J. Math. Anal., 27 (1996), pp. 1261–1269.

[12] W. Shen, *Traveling waves in time almost periodic structures governed by bistable nonlinearities* I *and* II, J. Differential Equations, to appear.

[13] K. W. Schaaf, *Asymptotic behavior and traveling wave solutions for parabolic functional-differential equations*, Trans. Amer. Math. Soc., 302 (1987), pp. 587–615.

[14] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.

[15] C. C. Travis and G. F. Webb, *Existence and stability for partial functional differential equations*, Trans. Amer. Math. Soc., 200 (1974), pp. 395–418.

[16] A. I. Volpert, Vitaly A. Volpert, and Vladimir A. Volpert, *Traveling Wave Solutions of Parabolic Systems*, Transl. Math. Monogr. 140, AMS, Providence, RI, 1994.

[17] J. Wu, *Theory and Applications of Partial Functional-Differential Equations*, Appl. Math. Sci. 119, Springer-Verlag, New York, 1996.

[18] X. Zou and J. Wu, *Existence of traveling wave fronts in delayed reaction-diffusion systems via the monotone iteration method*, Proc. Amer. Math. Soc., 125 (1997), pp. 2589–2598.

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 535–560

QUALITATIVE BEHAVIOR OF SOLUTIONS OF CHEMOTACTIC DIFFUSION SYSTEMS: EFFECTS OF MOTILITY AND CHEMOTAXIS AND DYNAMICS*

XUEFENG WANG†

Abstract. Chemotaxis is the oriented movement of cells in response to the concentration gradient of chemical substances in their environment. We consider the situation of a single bacterial population which responds chemotactically to a nutrient diffusing from an adjacent phase not accessible to the bacteria. The concentration and density of the substrate and cells (resp.) satisfy a quasi-linear parabolic system with nonlinear boundary condition. Our first set of results addresses the effects of two important biological parameters $\lambda > 0$ and $\chi > 0$ on the steady states, where λ measures the (random) motility of bacteria and χ the magnitude of chemotactic response (or sensitivity) to the chemical. Our second set of results concerns the global boundedness of time-dependent solutions and stability issues of trivial and nontrivial steady states with small amplitudes.

Key words. chemotaxis, motility, monotonicity, Helly's theorem, concentration, blowup, bifurcation, existence and uniqueness, global boundedness, stability, decay rates, Moser–Alikakos iteration

AMS subject classifications. 92B05, 35B25, 35B32, 35B35, 35B40, 35K57

PII. S0036141098339897

1. Introduction and description of main results.

Chemotaxis is the oriented movement of cells in response to the concentration gradient of chemical substances in their environment. Motility, on the other hand, refers to the random diffusivity of cells. It is known that chemotaxis and motility can affect not only the distribution but also the growth of cells. Sometimes they can even be a decisive advantage for a species of cells in its competition for a limited resource with another species of less chemotactic or more motile cells which has even superior growth kinetics [LAK], [PW].

Keller and Segal [KS] seem to have been the first to propose mathematical models for chemotaxis. To illustrate the general idea in modeling chemotaxis, we consider a species of cells which responds chemotactically to a chemical. Let $u(x, t)$ and $v(x, t)$ be the concentration and density of the chemical and cells, respectively. If the chemical is diffusive, then by Fick's law the random diffusive flux is given by $-D_1\nabla u$, where $D_1 > 0$ is assumed to be a constant. The cell flux is assumed to be the combination of the random diffusive flux and chemotactic flux, with the latter parallel to ∇u, so the cell flux takes the form $-D_2\nabla v + \chi v\phi'(u)\nabla u$, where $D_2 > 0$ and χ are constants, and $\phi'(u) > 0$. D_2 is the *motility* of cells and measures the ability of cells to diffuse randomly; χ (positive if the chemical is an attractant and negative if it is a repellent) is called the *chemotaxis coefficient* and measures the magnitude of response of cells to the chemical. $\phi(u)$ is called the *sensitivity function*. We need the factor $\phi'(u)$ in the chemotaxis flux to reflect the fact that the sensitivity of cells to the chemical may vary at different levels of chemical concentration.

Now by the law of conservation of mass, we are led to the following quasi-linear

*Received by the editors June 8, 1998; accepted for publication (in revised form) May 5, 1999; published electronically February 18, 2000. This work was partially supported by NSF DMS-9622867.
http://www.siam.org/journals/sima/31-3/33989.html

†Department of Mathematics, Tulane University, New Orleans, LA 70118 (xdw@math.tulane.edu).

system

$$\begin{cases} u_t = D_1 \Delta u + h(u,v), \\ v_t = \nabla \cdot (D_2 \nabla v - \chi v \nabla \phi(u)) + k(u,v), \end{cases} \tag{1.1}$$

where $h(u,v)$ is the creation-degradation rate of the chemical and $k(u,v)$ is the birth-death rate of cells.

Physiologists are interested in the effects of cell motility and chemotaxis on the population growth. To elucidate such effects, Lauffenburger, Aris, and Keller [LAK] investigated the situation of a single bacterial population in a one-dimensional medium with finite length, with growth limited by a nutrient diffusing from an adjacent phase not accessible to the bacteria. Their mathematical model is the following (nondimensionalized form):

$$\begin{cases} u_t = u_{xx} - f(u)v, \ 0 \le x \le 1, \ t > 0, \\ v_t = (\lambda v_x - \chi v(\phi(u))_x)_x + (kf(u) - \theta)v, \\ u_x(0) = 0, \ u_x(1) = h(1 - u(1)), \\ \lambda v_x - \chi v(\phi(u))_x = 0 \quad \text{at} \quad x = 0, 1. \end{cases} \tag{1.2}$$

Here λ, χ, k, θ, and h are positive constants except that χ is allowed to equal zero; u is the concentration of the nutrient and v the density of the bacteria; $f(u)$ is the consumption rate of the nutrient per cell; the term $(kf(u) - \theta)v$ in the v-equation represents that the bacteria have a *Malthusian* (or exponential) growth with $kf(u)$ and θ measuring the birth and death rate of cells. The boundary condition for u at $x = 1$ reflects the assumption that the flux of u at $x = 1$ is proportional to the difference of chemical concentration on two sides of the point $x = 1$ with h being the proportionality constant. The boundary condition for v says that there is no flux for the cells at the boundary.

From biological and technical considerations, we require f and ϕ satisfying

$$\begin{gathered} f(0) = 0, \ f'(u) > 0, \quad \text{and} \quad \phi'(u) > 0 \quad \text{for} \quad u \in [0, \infty), \\ f \in C^3([0,\infty)) \quad \text{and} \quad \phi \in C^5([0,\infty)). \end{gathered} \tag{1.3}$$

Typical choices for f and ϕ are $f(u) = au/(b+u)$, $\phi(u) = u$, $\phi(u) = \log(c+u)$, $\phi(u) = u/(1+cu)$, etc. In [LAK], $\phi(u)$ is taken to be u.

Numerical simulations on *steady states* of (1.2) (with $\phi(u) = u$, χ proportional to λ) led the authors of [LAK] to the following interesting observations.

(C1) Random motility λ may lead to decreased population (at least in nonchemotactic case $\chi = 0$).

(C2) Chemotaxis coefficient χ acts to increase population size.

These are the primary motivations of the current paper. We shall

(i) study the effects of large or small λ or χ on positive steady states of (1.2);

(ii) do so when the bacteria have a different growth type (logistic growth, which was not considered in [LAK]);

(iii) study the important related dynamical issues, such as stability of steady states and boundedness of global solutions.

The results obtained do lead to new phenomena and new understandings beyond (C1), (C2), and the numerical simulations in [LAK], though (C1) and (C2) are not "proved" completely.

Before describing our main results, we mention that in [Z] (which to our knowledge is the only previous rigorous work on (1.2)), Zeng proved that (i) if $\theta \ge kf(1)$, the only steady state of (1.2) is the trivial one: $(u, v) = (1, 0)$; (ii) if $0 < \theta < kf(1)$, then (1.2) has a positive steady state (u, v); (iii) any positive steady state (u, v) of (1.2) satisfies $0 < u < 1$, $u' > 0$, and $v' > 0$. (In [Z], $\phi(u) = u$ is assumed; but the arguments there go through for general ϕ satisfying (1.3).)

In this paper, concerning positive steady states of (1.2), we obtain what can be roughly described as follows.

(R1) Small motility λ and large chemotaxis coefficient χ (when compared to λ, that is, χ is large and λ/χ is small) have the same effect on the distribution and the total population of the bacteria: v concentrates and is approximately an δ-function at the boundary point $x = 1$; moreover, the total population of bacteria $\int_0^1 v(x)dx$ is close to $kh(1-c)/\theta$, u is close to c, where c is the constant in $(0, 1)$ such that $kf(c) = \theta$.

(R2) Small motility λ and large chemotaxis coefficient χ (when compared to λ) act to increase the total population of bacteria: $\int_0^1 v(x)dx$ is maximized in the limit $\lambda \to 0$ or $\lambda/\chi \to 0$ and $\chi \to \infty$.

(R3) When both λ and χ are large and λ is at least at the order of χ, v no longer concentrates at $x = 1$; in the limit, both u and v converge in $C^1[0, 1]$ norm to functions which, in the case $\lambda/\chi \to \infty$, can be uniquely determined.

(For the precise statements of the results described above, see Theorem 2.1.)

(R1) and (R3) were not observed numerically or by the formal arguments in [LAK] or in any other places, to the best of our knowledge. According to (C1) and (C2), the functions

$$\lambda \to \int_0^1 v_\lambda(x)dx, \quad \chi \to \int_0^1 v_\chi(x)dx \tag{1.4}$$

(subscripts indicating the dependence of v on the parameters, *not* partial derivatives) should be monotone decreasing and increasing (resp.), at least when χ is proportional to λ. It seems that this kind of question has not been rigorously studied for the classical reaction-diffusion systems (nonchemotactic). In fact, for many of them, even the uniqueness or multiplicity problem has largely remained open. For our system (1.2) with $\chi = 0$, the uniqueness of positive steady state was pointed out by Yuan Lou and later by Junping Shi. If $\chi > 0$, we are unable to prove that the functions in (1.4) are single-valued.

We have seen from (R1) that for small λ or large χ (compared to λ), the bacteria concentrate in a small neighborhood of the boundary point $x = 1$, which is a "fertile zone", because the nutrient diffuses into the medium through $x = 1$. This should promote the growth of bacteria if they have a Malthusian growth, as assumed in the model (1.2). This may not promote the growth of bacteria if their growth type is logistic, that is, if large density of bacteria leads to negative growth due to overcrowding.

Motivated by this, we propose the following model which is a slight modification of (1.2).

$$\begin{cases} u_t = u_{xx} - f(u)v, \ 0 \le x \le 1, \ t > 0, \\ v_t = (\lambda v_x - \chi v(\phi(u))_x)_x + (kf(u) - \theta - \beta v)v, \\ u_x(0) = 0, \ u_x(1) = h(1 - u(1)), \\ \lambda v_x - \chi v(\phi(u))_x = 0 \ \text{ at } \ x = 0, 1, \end{cases} \tag{1.5}$$

where β is a positive constant. At the constant level of the chemical concentration u, the bacteria have logistic growth.

We shall show that as in the Malthusian case, (1.5) has a positive steady state if and only if $0 < \theta < kf(1)$ (see Theorem 3.1). Our results on qualitative behavior of positive steady states can be described as follows.

(R4) In sharp contrast to the Malthusian case, large chemotaxis coefficient χ (compared to λ) is *detrimental* to the growth of bacteria: u is close to 1 throughout the interval $[0, 1]$, v concentrates at the boundary point, but the total population $\int_0^1 v(x)dx$ is close to zero.

(R5) If $\chi/\sqrt{\lambda}$ remains bounded as $\lambda \to 0$, then v does not concentrate at point $x = 1$ and $\liminf_{\lambda \to 0} \int_0^1 v_\lambda(x)dx > 0$; these are true if χ is only bounded as $\lambda \to 0$, in the special case $\phi(u) = u$.

(R6) The conclusion in (R3) holds in the logisitc case (1.5).

See Theorem 3.3 for precise statements. Notice that in the logisitc case, small motility λ and large chemotaxis coefficient χ no longer have the same effect on v and the total population $\int_0^1 v(x)dx$, in contrast to the Malthusian case. We conjecture that in the logisitc case, small motility λ is also detrimental to the total population $\int_0^1 v(x)dx$, that is, in the limit $\lambda \to 0$, $\int_0^1 v_\lambda(x)dx$ is minimized (assuming that χ remains bounded).

We remark that the purpose of the logistic kinetics we use here is not to prevent the blowup (in finite time or infinite time) of v (concentration) or $\int_0^1 v(x,t)dx$ (total population) because both remain bounded as $t \to \infty$ even in the Malthusian case (see (R7)). Instead, it is used to model the inhibition on growth due to the competition for the nutrient among cells at locations of aggregation.

All of the above results concern the behavior of positive steady states of (1.2) and (1.5). On the dynamical issues for (1.5) ($\beta = 0$ allowed so that the Malthusian case is included), we have the following.

(R7) (1.5) has a unique bounded global (in time) solution.

(R8) When $\theta \geq kf(1)$, the trivial steady state $(1, 0)$ is globally asymptotically stable; when $0 < \theta < kf(1)$, it is unstable.

(R9) For θ less than but close to $kf(1)$, the positive steady state of (1.5) is unique and asymptotically stable. See Theorems 4.8, 5.1, and 5.2 for precise statements.

Recall that (1.5) has a positive steady state if and only if $0 < \theta < kf(1)$. We conjecture that for the full parameter range $\theta \in (0, kf(1))$, the positive steady state of (1.5) is unique and globally asymptotically stable. If this is true, it would imply that the behavior of the positive steady state represents that of the time-dependent solution for large time. The results (R7)–(R9) support this conjecture. It has been an outstanding open problem for many classical (nonchemotactic) diffusion systems to obtain uniqueness and stability of steady states for the full range of parameters. For our problems, we have one more significant difficulty to overcome: the nonlinear boundary condition for v. Indeed, in obtaining even the stability in (R9) for θ close to $kf(1)$ by using the bifurcation method, we cannot directly apply the standard result of Crandall–Rabinowitz because the linearized differential operators have varying domains, due to the nonlinear boundary condition.

The paper is organized as follows. We study (i) the behavior of steady states in the Malthusian case in section 2; (ii) existence and behavior of steady states in the logisitc case in section 3; (iii) global existence and boundedness of time-dependent solutions in section 4; (iv) stability of steady states in section 5.

We remind the reader that we assume the condition (1.3) throughout this paper.

2. Behavior of steady states in Malthusian case. We first write down the system satisfied by steady states of (1.2):

$$\begin{cases} u'' = f(u)v, \ x \in [0,1], \\ (\lambda v' - \chi v(\phi(u)')') + (kf(u) - \theta)v = 0, \\ u'(0) = 0, \ u'(1) = h(1 - u(1)), \\ \lambda v' - \chi v(\phi(u))' = 0 \ \text{ at } \ x = 0, 1. \end{cases} \tag{2.1}$$

Recall [Z] that (2.1) has a positive steady state if and only if $0 < \theta < kf(1)$. We shall assume this throughout this section.

The purpose of this section is to prove the following theorem.

THEOREM 2.1. *Let* (u_i, v_i) *be positive solutions of* (2.1) *with* $(\lambda, \chi) = (\lambda_i, \chi_i)$.

(i) *If* $\lambda_i \to 0$, *then* $u_i \to$ *constant* c *uniformly on* $[0,1]$, *where* $kf(c) = \theta$, v_i *concentrates and blows up at* $x = 1$, *i.e.,* v_i *converges to zero uniformly outside any left neighborhood of* $x = 1$ *and* $v_i(1) \to \infty$; *moreover, the total population of bacteria* $\int_0^1 v_i(x) \to kh(1-c)/\theta$. *(Thus* v_i *converges to a constant multiple of the* δ*-function centered at* $x = 1$.*)*

(ii) *If* $\lambda_i/\chi_i \to 0$ *as* $\chi_i \to \infty$ *or as* $\lambda_i \to \infty$ *(so* χ_i *is relatively large), then the conclusion in* (i) *is true.*

(iii) *Suppose that* $\lim_{\lambda_i \to \infty} \lambda_i/\chi_i = a$ *exists and* $a \in (0, \infty]$. *Then after passing to a subsequence,* $u_i(x) \to u_\infty(x)$, $v_i(x) \to v_\infty(x)$ *in* $C^1([0,1])$, *where* u_∞ *and* v_∞ *satisfy*

$$\begin{cases} u''_\infty = f(u_\infty)v_\infty, \ u_\infty > 0, \ v_\infty > 0, \ x \in [0,1], \\ u'_\infty(0) = 0, \ u'_\infty(1) = h(1 - u_\infty(1)), \\ \textit{constraint: } \int_0^1 (kf(u_\infty(x)) - \theta)v_\infty dx = 0. \end{cases} \tag{2.2}$$

Moreover, if $a = \infty$, *then* v_∞ *is a constant, and* (u_∞, v_∞) *is uniquely determined by* (2.2) *and hence the whole sequence* (u_i, v_i) *converges to* (u_∞, v_∞); *if* a *is finite, then* $v_\infty = (\text{const}M)\exp(\phi(u_\infty)/a)$.

Proof of (i). *Step* 1. By integrating both sides of the v-equation in (2.1), we see

$$\int_0^1 (kf(u_i(x)) - \theta)v_i(x)dx = 0. \tag{2.3}$$

Step 2. We show that for any $x \in [0,1)$, $\limsup_{i\to\infty} v_i(x) < \infty$.

If this is not true, then there exists $x_0 \in [0,1)$ and a sequence $i \to \infty$ such that $\lim_{i\to\infty} v_i(x_0) = \infty$. Since $u''_i > 0$ and by the boundary condition on u_i, we have $0 \le u'_i(x) \le u'_i(1) < h$ for $x \in [0,1]$. Then by virtue of the Azela–Ascoli theorem, we obtain that, after passing to a subsequence,

$$u_i(x) \to u_0(x) \ \text{ uniformly on } \ [0,1] \ \text{ as } \ i \to \infty. \tag{2.4}$$

On the other hand, by integrating the u-equation, we obtain

$$\begin{aligned} h > u'_i(1) - u'_i(0) &= \int_0^1 f(u_i(x))v_i(x)dx \\ &\ge \int_{x_0}^1 f(u_i(x))v_i(x)dx \ge \int_{x_0}^1 f(u_i(x))v_i(x_0)dx \end{aligned}$$

(recall v_i is increasing on $[0,1]$), which in turn implies $h \ge \int_{x_0}^1 f(u_0(x))\infty\, dx$ by Fatou's lemma. Thus $u_0 \equiv 0$ on $[0,1]$. But this and (2.4) imply that the integral in (2.3) is negative for i large enough; thus, we have a contradiction!

Step 3. By Step 2, the monotonicity of v_i and Helly's theorem, we have that after passing to yet another subsequence,

$$v_i(x) \to \text{some } v_0(x) \text{ pointwise on } [0,1) \text{ as } i \to \infty, \tag{2.5}$$

where v_0 is, of course, nondecreasing on $[0,1)$. Since the case when χ_i becomes unbounded as $\lambda_i \to 0$ can be covered by part (ii) of Theorem 2.1, here we assume that χ_i remains bounded as $\lambda_i \to 0$. After passing to a subsequence, we assume $\chi_i \to$ some $\chi_0 \geq 0$. We claim

$$\int_0^x (kf(u_0(y)) - \theta)v_0(y)dy = \chi_0 v_0(x)u_0'(x)\phi'(u_0(x)) \text{ almost everywhere (a.e.) on } [0,1). \tag{2.6}$$

To see this, we integrate the v-equation in (2.1) over the interval $[0,x]$ to obtain

$$\lambda_i v_i'(x) - \chi_i v_i(x)u_i'(x)\phi'(u_i(x)) + F_i(x) = 0, \; x \in [0,1], \tag{2.7}$$

where $F_i(x) = \int_0^x (kf(u_i(y)) - \theta)v_i(y)dy$. Observe that by Step 2 and Lebesgue's dominated convergence theorem, we have that $F_i(x) \to F_0(x) \equiv \int_0^x (kf(u_0(y)) - \theta)v_0(y)dy$ pointwise on $[0,1)$ as $i \to \infty$. Integrating (2.7) from 0 to x and sending λ_i to 0, we have

$$\int_0^x [-\chi_0 v_0(y)u_0'(y)\phi'(u_0(y)) + F_0(y)]dy = 0, \; x \in [0,1).$$

Now (2.6) follows.

Step 4. We show $v_0 \equiv 0$ on $[0,1)$ and $u_0 \equiv$ const c with $kf(c) = \theta$. There are two cases to consider.

Case 1. $\chi_0 = 0$. In light of (2.6), we have

$$(kf(u_0(x)) - \theta)v_0(x) = 0, \text{ a.e. on } (0,1). \tag{2.8}$$

If $v_0 \not\equiv 0$ on $[0,1)$, then there exists $x_0 \in [0,1)$ such that $v_0(x) > 0$ for $x \in [x_0,1)$. Then (2.8) implies that $kf(u_0(x)) = \theta$, i.e., $u_0 \equiv c$ on $[x_0,1]$. On the other hand, by integrating the u-equation, it is easy to see

$$u_0'(x) = \int_0^x f(u_0(y))v_0(y)dy, \; x \in [0,1). \tag{2.9}$$

Now since u_0 is a constant on $[x_0,1]$, $\int_0^x f(u_0(y))v_0(y)dy \equiv 0$ and hence $v_0 \equiv 0$ on $[0,1)$, contradicting the assumption that $v_0(x) > 0$ for $x \in [x_0,1)$. Thus $v_0 \equiv 0$ on $[0,1)$ and by (2.9), u_0 is a constant.

By (2.3), $f(u_i(1)) > \theta/k$. Thus $f(u_0) \geq \theta/k$. If $f(u_0) > \theta/k$, then the integral in (2.3) is negative for small enough λ_i, which is impossible. Therefore $kf(u_0) = \theta$. Step 4 is finished in the case where $\chi_0 = 0$.

Case 2. $\chi_0 > 0$. If we can show $v_0 \equiv 0$ on $[0,1)$, then the rest of the arguments are exactly the same as in Case 1. Suppose there exists $x_0 \in [0,1)$ such that $v_0 > 0$ on $[x_0,1)$. We want to show that $u_0' \equiv 0$ on $[x_0,1)$. Otherwise, there exists $x_1 \in (x_0,1)$ such that $u_0'(x_1) > 0$. Then by (2.6), we have $\int_0^{x_1} (kf(u_0(y)) - \theta)v_0(y)dy > 0$; hence, $kf(u_0(x_1)) > \theta$. Therefore, $kf(u_i(x)) - \theta > 0$ for $x \in [x_1,1)$ and large enough i. Consequently, for such i, $\int_{x_1}^1 (kf(u_i(y)) - \theta)v_i(y)dy > 0$. Now observe

$$\begin{aligned} 0 = \lim_{i\to\infty} \int_0^1 (kf(u_i(y)) - \theta)v_i(y)dy &= \lim_{i\to\infty} \left(\int_0^{x_1} + \int_{x_1}^1 \right) (kf(u_i(y)) - \theta)v_i(y)dy \\ \geq \lim_{i\to\infty} \int_0^{x_1} (kf(u_i(y)) - \theta)v_i(y)dy &= \int_0^{x_1} (kf(u_0(y)) - \theta)v_0(y)dy > 0; \end{aligned}$$

we have a contradiction! We have shown that $u_0' \equiv 0$ on $[x_0, 1)$ under the assumption that $v_0 > 0$ on $[x_0, 1)$. Now as in Case 1, we are led to a contradiction.

Step 5. We now only need to show $\int_0^1 v_i(x)dx \to kh(1-c)/\theta$. By (2.3),

$$\theta \lim_{i\to\infty} \int_0^1 v_i(y)dy = k \lim_{i\to\infty} \int_0^1 f(u_i(y))v_i(y)dy = k \lim_{i\to\infty} (u_i'(1) - u_i'(0))$$
$$= kh \lim_{i\to\infty} (1 - u_i(1)) = kh(1-c).$$

This completes the proof of (i) of Theorem 2.1.

Proof of (ii) *of Theorem* 2.1. The proof is basically the same as that for part (i). The revisions needed are to replace (2.6) by

$$v_0(x)u_0'(x)\phi'(u_0(x)) = 0, \ x \in [0, 1), \tag{2.10}$$

then to change the proof in Step 4 slightly. To show (2.10), we divide both sides of (2.7) by χ_i and then send χ_i or λ_i to infinity. Combining (2.9) with (2.10), we immediately see that $v_0 \equiv 0$ and $u_0 \equiv$ constant on $[0, 1)$. Now using the arguments in Step 4, we have $u_0 \equiv c$ on $[0, 1]$. (ii) of Theorem 2.1 is proved.

Before proving part (iii) of Theorem 2.1, we prove the following.

LEMMA 2.2. *If λ is bounded away from* 0 *and if χ/λ is bounded, then v is bounded on* $[0, 1]$. *Here λ and χ are allowed to vary, dependently or independently on each other.*

Proof. From (2.7) (dropping the subscripts), we have

$$v'(x) - \frac{\chi u'(x)}{\lambda} v(x)\phi'(u(x)) \le \theta \int_0^1 v(x)dx/\lambda.$$

By (2.3) and the u-equation in (2.1), $\theta \int_0^1 v(x)dx = ku'(1) \le hk$. Using this and the fact that $u'(x) \le u'(1) \le h$, we see that

$$v'(x) - \frac{\chi h}{\lambda} v(x)\|\phi'\|_{L^\infty} \le hk/\lambda.$$

From this it follows that

$$v(x) \le v(0)e^{\frac{\chi h\|\phi'\|_{L^\infty}}{\lambda}x} + \frac{k}{\chi\|\phi'\|_{L^\infty}}(e^{\frac{\chi h\|\phi'\|_{L^\infty}}{\lambda}x} - 1), \ x \in [0, 1],$$

where the second term should be understood as equal to its limit when $\chi = 0$. On the other hand, as argued in the proof of (i) of Theorem 2.1, $v(0)$ must remain bounded. Thus v remains bounded on $[0, 1]$. Lemma 2.2 is proved.

Proof of (iii) *of Theorem* 2.1. We shall consider only the case $\chi_i \to \infty$ with $\lambda_i/\chi_i \to \infty$. The other case can be handled in the same way.

By the u-equation in (2.1) and Lemma 2.2, u_i'' is bounded on $[0, 1]$. Thus by the Azela–Ascoli theorem, $u_i \to$ some u_∞ in $C^1([0, 1])$, as $i \to \infty$, after passing to a subsequence. From (2.7) and Lemma 2.2, we easily see that $v_i' \to 0$ uniformly on $[0, 1]$ as $i \to \infty$. This implies that after passing to a subsequence, $v_i \to$ const v_∞ in $C^1([0, 1])$ as $i \to \infty$. We now show that $v_\infty > 0$. Otherwise, $u_\infty' \equiv 0$ on $[0, 1]$ (see (2.9)) and then $u_\infty \equiv 1$ on $[0, 1]$. Then for i large, the integral in (2.3) is positive (recall $0 < \theta < kf(1)$). Impossible.

It is easy to see that (u_∞, v_∞) satisfies (2.2). To show the uniqueness of (u_∞, v_∞), we first observe that by the comparison principle, for a fixed $v_\infty \ge 0$, (2.2) without the constraint has at most one solution u_∞ (here we need the assumption $f'(u) > 0$). Thus it suffices to show that v_∞ is unique. By the comparison principle again, we have that u_∞ is monotonically decreasing with respect to constant v_∞. Thus there exists at most one $v_\infty > 0$ such that $\int_0^1 kf(u_\infty(x))dx = \theta$, i.e., the constraint in (2.2) is satisfied.

Remark 2.3. Simple calculations below show that in the cases considered in (i) and (ii) of Theorem 2.1, $\int_0^1 v(x)dx$ is maximized in the limit $\lambda \to 0$. By (2.3), we have $\int_0^1 v(x)dx = \frac{k}{\theta}\int_0^1 f(u(x))v(x)dx = \frac{k}{\theta}u'(1) = \frac{kh}{\theta}(1-u(1))$. By (2.3) again, $u(1) > c$. Thus $\lim_{\lambda\to 0}\int_0^1 v(x)dx = \frac{kh}{\theta}(1-c) > \int_0^1 v(x)dx$. The same is true when $\chi \to \infty$ and $\lambda/\chi \to 0$.

3. Positive steady states in the logistic case.

3.1. Existence. The steady states of (1.5) satisfy

$$\begin{cases} u'' = f(u)v, \\ \lambda v'' - \chi(v\phi'(u)u')' + (kf(u) - \theta - \beta v)v = 0, \\ u'(0) = 0,\ u'(1) = h(1-u(1)), \\ \lambda v' - \chi v\phi'(u)u' = 0 \quad \text{at} \quad x = 0, 1, \end{cases} \tag{3.1}$$

where $\beta > 0$ is a constant.

Theorem 3.1. (i) *For $\theta \geq kf(1)$,* (3.1) *has no positive solutions.*

(ii) *For $0 < \theta < kf(1)$,* (3.1) *has a positive solution.*

(iii) *For θ less than but close to $kf(1)$, the positive solution of* (3.1) *is unique.*

The proof of (i) is easy. To prove (ii), we can use either the decoupling method combined with bifurcation techniques (as in [Z] for (3.1) with $\beta = 0$), or the bifurcation method directly applied to the system (as developed in [BB] for another system). We choose the second approach, which also prepares results and notation for our stability analysis in section 5. The proof is, in spirit, similar to that in [Z] and will be only sketched here.

The following can be easily observed by inspecting (3.1) and is useful in proving (i).

Lemma 3.2. *If (u, v) is a positive solution of* (3.1), *then $u(x) < 1$ for $0 \leq x \leq 1$ and $u'(x) > 0$ for $0 < x \leq 1$.*

Proof of (i) *of Theorem* 3.1. Integrating the v-equation in (3.1) and using the boundary condition, we have

$$\int_0^1 (kf(u) - \theta - \beta v)v\, dx = 0. \tag{3.2}$$

If $\theta \geq kf(1)$ and v is positive, then the integrand is strictly negative because $u < 1$ and $f'(u) > 0$.

Proof of (ii) *of Theorem* 3.1. We first convert (3.1) into a bifurcation problem with $\mu \equiv kf(1) - \theta$ as a parameter. Let $w = 1 - u$. Then (3.1) is transformed into

$$\begin{cases} -w'' = f(1-w)v, \\ \lambda v'' + \chi(v\phi'(1-w))'w' - \chi v\phi'(1-w)f(1-w)v + (kf(1-w) - \theta - \beta v)v = 0, \\ w'(0) = 0 = w'(1) + hw(1), \\ \lambda v' + \chi v\phi'(1-w)w' = 0 \quad \text{at} \quad x = 0, 1. \end{cases} \tag{3.3}$$

We seek a positive solution (w, v) of (3.3) with $1 - w > 0$. For technical reasons, we need to extend f and ϕ on $\mathbb{R}$. By the regularity assumption (1.3) for f and ϕ, we can extend them on the whole real line so that f and ϕ have the same regularity on $\mathbb{R}$ as in (1.3). This extension will not affect the set of positive solutions of (3.3) with $w < 1$.

Observe that (3.3) is equivalent to the following functional equation:

$$\begin{pmatrix} w \\ v \end{pmatrix} = \begin{pmatrix} 0 & f(1)K_1 \\ 0 & K_2 \end{pmatrix} \begin{pmatrix} w \\ v \end{pmatrix} + \mu \begin{pmatrix} 0 & 0 \\ 0 & K_2 \end{pmatrix} \begin{pmatrix} w \\ v \end{pmatrix} + \begin{pmatrix} R_1(w,v) \\ R_2(w,v) \end{pmatrix}, \tag{3.4}$$

where K_1 is the inverse of $-\frac{d^2}{dx^2}$ with the w-boundary condition in (3.3), K_2 is the inverse of $-\lambda\frac{d^2}{dx^2}+1$ with the homogeneous Neumann boundary condition, $R_1(w,v) = K_1[(f(1-w)-f(1))v]$, $B(w,v) = -w'(1)v(1)\phi'(1-w(1))[\chi x^2/2\lambda - K_2(-\chi+\chi x^2/2\lambda)]$,

$$R_2(w,v) = K_2[\chi(v\phi'(1-w))'w' - \chi v\phi'(1-w)f(1-w)v + (kf(1-w)-kf(1)-\beta v)v] + B(w,v).$$

By the elliptic regularity theory, the linear operators K_1 and K_2 map $C^\alpha([0,1])$ into $X \equiv C^{\alpha+1}([0,1])$ and are compact. Denote by K the first matrix in (3.4), by L the second matrix, and by $R(w,v)$ the R_1, R_2 vector. Then (3.4) can be rewritten as

$$(I - K - \mu L)\begin{pmatrix} w \\ v \end{pmatrix} - R(w,v) = 0, \ (w,v) \in X \times X. \tag{3.5}$$

Denote the left-hand side of (3.5) by $F(\mu,(w,v))$. Obviously $F(\mu,(0,0)) = 0$. We show that a local bifurcation occurs at $(\mu,(w,v)) = (0,(0,0))$ by using the Crandall–Rabinowitz bifurcation theorem (see Lemma 1.1 in [CR]). To this end, we need to show (details omitted)

(a) $F : \mathbb{R} \times X \times X \to X \times X$ is C^2 smooth;

(b) $\dim N(F_{(w,v)}(0,(0,0))) = 1 = \operatorname{codim} R(F_{(w,v)}(0,(0,0)))$;

(c) $F_{\mu(w,v)}(0,(0,0))(\bar{w}_0,\bar{v}_0) \notin R(F_{(w,v)}(0,(0,0)))$, where $(\bar{w}_0,\bar{v}_0) = (K_1 f(1), 1)$ spans $N(F_{(w,v)}(0,(0,0)))$.

Now let Z be any complement of span $\binom{\bar{w}_0}{\bar{v}_0}$ in $X \times X$. Then by the Crandall–Rabinowitz theorem, there exist a positive ε and C^1 smooth functions $\mu : (-\varepsilon,\varepsilon) \to \mathbb{R}$, $\binom{\psi_1}{\psi_2} : (-\varepsilon,\varepsilon) \to Z$ such that $\mu(0) = 0$, $\psi_1(0) = 0 = \psi_2(0)$, and $F(\mu(s),(w(s),v(s))) = 0$, where $w(s) = s\bar{w}_0 + s\psi_1(s)$, $v(s) = s\bar{v}_0 + s\psi_2(s)$, $s \in (-\varepsilon,\varepsilon)$. Moreover, all solutions $(\lambda,(w,v))$ of $F = 0$ near $(0,(0,0))$ are either on $(w,v) = 0$ or on the curve $(\mu(s),(w(s),v(s)))$, $s \in (-\varepsilon,\varepsilon)$. Notice that for $s > 0$ small, $w(s)$ and $v(s)$ are positive functions (of x) on $[0,1]$ because so are $\bar{w}_0$ and $\bar{v}_0$.

Let S be the closure of the set of solutions $(\mu,(w,v)) \in \mathbb{R} \times X \times X$ of $F = 0$ with $(w,v) \neq (0,0)$. Let C be the maximal subcontinuum of S passing through $(0,(0,0))$. Let C^+ be the maximal subcontinuum of the closure of $C - \{(\mu(s),(w(s),v(s))) \mid -\varepsilon < s < 0\}$ which meets $(0,(0,0))$. Then by combining the reflection argument in [R, Theorem 1.27] and [BB, Theorem 3.2] we have that C^+ either meets "infinity" or meets $(\widehat{\mu},(0,0))$ where $\widehat{\mu} \neq 0$ and $I - K - \widehat{\mu}L$ is not invertible, or C^+ contains a pair of points $(\mu,(w,v))$ and $(\mu,-(w,v))$. We proceed to show $C^+ \backslash \{(0,(0,0))\} \subset \mathbb{R} \times X^+ \times X^+$, where X^+ is the positive cone of X consisting of functions positive on $[0,1]$. We omit the details.

Now we see that C^+ cannot possibly satisfy either of the last two alternatives mentioned above and hence C^+ meets "infinity."

By an argument similar to the one for $C^+ \backslash \{(0,(0,0))\} \subset \mathbb{R} \times X^+ \times X^+$, we can show that if $(\mu,(w,v)) \in C^+ \backslash \{(0,(0,0))\}$, then $w < 1$; and by integrating both sides of v-equation in (3.1) on $[0,1]$, we see $\mu > 0$ (thus the bifurcation curve "turns to the right" at $(0,(0,0))$).

We now show that the projection of $C^+\backslash\{(0,(0,0))\}$ onto the μ-axis contains the interval $(0, kf(1))$. It is sufficient to obtain an a priori bound for (w, v) if $(\mu, (w, v))$ in $\mathbb{R} \times X^+ \times X^+$ is a solution of $F = 0$ and if $\mu \in (0, kf(1))$. We proceed to obtain such a bound for v first. By (3.2), we have

$$\beta \int_0^1 v^2 dx < \mu \int_0^1 v\, dx < kf(1) \left(\int_0^1 v^2 dx \right)^{1/2}, \tag{3.6}$$

and hence

$$\int_0^1 v(x) dx \leq \left(\int_0^1 v^2 dx \right)^{1/2} < \frac{kf(1)}{\beta}. \tag{3.7}$$

Now integrating the v-equation in (3.1) from 0 to x, and using (3.7), we obtain

$$\lambda v'(x) + \chi \phi'(1-w) w' v = g(x), \ x \in [0,1],$$

where $|g|$ is bounded by a constant depending only on k and β. It follows that

$$v(x) = e^{\frac{\chi}{\lambda}\phi(1-w(x))}(v(0)e^{-\chi\phi(1-w(0))/\lambda} + \frac{1}{\lambda}\int_0^x g(x) e^{-\chi\phi(1-w(x))/\lambda} dx). \tag{3.8}$$

Thus if $v(0)$ remains bounded for $\mu \in (0, kf(1))$, so does $\|v\|_{L^\infty}$ (recall $0 < w < 1$). If $v(0)$ does not remain bounded for $\mu \in (0, kf(1))$, then there exists a sequence $(\mu_n, (w_n, v_n)) \in (0, kf(1)) \times X^+ \times X^+$ which are solutions of $F = 0$, with $\mu_n \to \mu_0 \in [0, kf(1)]$ and $v_n(0) \to \infty$. Then by (3.8), $v_n(x) \to \infty$ uniformly on $[0,1]$. This would contradict (3.2) for n large. Thus $\|v\|_{L^\infty}$ is bounded for $\mu \in (0, kf(1))$. Hence by (3.3), $\|w\|_{C^2}$ is bounded and consequently so is $\|v\|_{C^2}$ for $\mu \in (0, kf(1))$.

We have so far shown that for $\mu \in (0, kf(1))$, (3.3) always has a positive solution (w, v) with $w < 1$. Part (ii) of Theorem 3.1 is proved.

Proof of (iii) *of Theorem* 3.1. We use the notation in the above proof. We claim that as $\theta \to kf(1)$ from the left, that is, as $\mu \to 0^+$, $\|(w, v)\| \to 0$. Suppose otherwise; then there exists sequence $(\mu_n, (w_n, v_n)) \subset \mathbb{R}^+ \times X^+ \times X^+$ which are solutions of (3.5), with $\mu_n \to 0^+$ and $\|(w_n, v_n)\|$ bounded away from 0. By the last part of the above proof, $\|(w_n, v_n)\|_X$ is bounded. Then the compactness of K, L, and R implies that (w_n, v_n) has a subsequence converging to some nonzero (w_∞, v_∞), which satisfies (3.5) and hence (3.3) with $\theta = kf(1)$. By (3.2), this is impossible. The claim is proved.

Now by Step 2 in the proof of (ii), for θ less than but close to $kf(1)$, i.e., for μ positive but small, $(\mu, (w, v))$ is on the curve $(\mu(s), (w(s, \cdot), v(s, \cdot)))$, $0 < s < \varepsilon$. If we can show $\mu'(0) \neq 0$, then the desired uniqueness follows. By (3.2), we have

$$\mu(s) \int_0^1 v(s,x) dx = \int_0^1 k(f(1) - f(1 - w(s,x))) v(s,x) dx + \beta \int_0^1 v^2(s,x) dx.$$

Dividing both sides by s^2 and sending $s \to 0$, we obtain

$$\mu'(0) \int_0^1 \bar{v}_0 dx = \int_0^1 k \bar{v}_0 f'(1) \bar{w}_0 dx + \beta \int_0^1 \bar{v}_0^2 dx.$$

Recall $\bar{w}_0 = K_1 f(1)$ (which is positive) and $\bar{v}_0 = 1$. Thus $\mu'(0) > 0$. This completes the proof of (iii).

3.2. Behavior of positive steady states. The main result in this subsection is the following.

THEOREM 3.3. *Suppose* $0 < \theta < kf(1)$. *Let* (u_i, v_i) *be positive solutions of* (3.1) *with* $(\lambda, \chi) = (\lambda_i, \chi_i)$. *In* (i) *and* (ii) *below we assume* $\lim_{\chi_i \to \infty} \lambda_i/\chi_i = a \in [0, \infty]$ *exists.*

(i) *If* $a = 0$, *then* $u_i \to 1$ *in* $C^1([0,1])$, $v_i \to 0$ *uniformly outside any fixed neighborhood of* $x = 1$, $\int_0^1 v_i(x)dx \to 0$. *Moreover, for i large the maximum of* v_i *is achieved at* $x = 1$ *and* $v_i(1) \not\to 0$.

(ii) *If* $0 < a \leq \infty$, *then the conclusion in* (iii) *of Theorem* 2.1 *holds with the constraint in* (2.2) *replaced by*

$$\int_0^1 (kf(u_\infty) - \theta - \beta v_\infty)v_\infty dx = 0.$$

(iii) *Suppose that* $\lambda_i \to 0$ *and* $\chi_i/\sqrt{\lambda_i}$ *remains bounded. Then* $u_i \to u_0$ *in* $C^1([0,1])$, $v_i \to v_0$ *pointwise in* $[0,1)$, *where* u_0 *is the unique positive solution of*

$$\begin{cases} u_0''(x) = f(u_0)(kf(u_0) - \theta)/\beta, \ x \in (0,1), \\ u_0'(0) = 0, \ u_0'(1) = h(1 - u_0(1)), \end{cases} \tag{3.9}$$

satisfying $kf(u_0) - \theta \geq 0$, *and* $v_0 = (kf(u_0) - \theta)/\beta$ *is positive on* $[0,1]$ *(and hence* $\liminf_{i\to\infty} \int_0^1 v_i(x)dx > 0$*).*

(iv) *Suppose* $\phi(u) = u$. *If* $\chi_i \to 0$, $\lambda_i \to 0$ *as* $i \to \infty$, *then the conclusion in* (iii) *holds; moreover, the convergence* $v_i \to v_0$ *is uniform outside any fixed neighborhood of* $x = 1$. *If* $\chi_i \to$ *some positive constant* χ_0 *and* $\lambda_i \to 0$, *then after passing to a subsequence* $u_i \to u_0$ *in* $C^1([0,1])$, $v_i \to v_0$ *uniformly outside any fixed neighborhood of* $x = 1$, $v_i(1) \to \infty$, *where* v_0 *is a positive function continuous on* $[0,1]$, $kf(u_0) > \theta + \beta v_0$ *on* $[0,1)$.

REMARK 3.4. The boundedness condition for $\chi/\sqrt{\lambda}$ in (iii), we suspect, is only a technical one. We tend to believe that the condition $\chi \to 0$ as $\lambda \to 0$ is sufficient, as in the case $\phi(u) = u$. The difficulty arises when we do not know if v is monotone increasing for general ϕ.

Before we start the proof of the above theorem, we make some preliminary observations.

LEMMA 3.5. *Let* (u, v) *be a positive solution of* (3.1). *Then we have*

(i) $(\int_0^1 v^2(x)dx)^{1/2} \leq kf(1)/\beta$.

(ii) *Let* x_0 *be a local maximum of* v *in* $[0,1)$. *Then*

$$v(x_0) \leq \begin{cases} (kf(1) + \chi h^2 \|\phi''\|_{L^\infty([0,1])})/\beta; \\ k/(\chi \min_{[0,1]} \phi'(u)). \end{cases}$$

(iii) *If* χ/λ *is bounded, then* $v(1)$ *is bounded.*

Proof. (i) follows easily from integrating the v-equation of (3.1) over the interval $[0,1]$.

To show (ii), observe that v satisfies

$$\lambda v'' - \chi\phi''(u)(u')^2 v - \chi\phi'(u)v^2 f(u) - \chi\phi'(u)u'v' + (kf(u) - \theta - \beta v)v = 0. \tag{3.10}$$

At x_0, $v'' \leq 0$ and $v' = 0$ (if $x_0 = 0$, we first extend u and v evenly and we have that u and v are C^2 smooth on $[-1,1]$). Therefore,

$$v(x_0) < (kf(u(x_0)) - \chi\phi''(u(x_0))(u'(x_0))^2)/\beta.$$

This and the fact that $0 < u < 1,\ 0 < u' \le u'(1) < h$ imply the first inequality in (ii).

On the other hand, from the v-equation in (3.1), we deduce

$$\lambda v'(x) - \chi\phi'(u)u'v(x) + \int_0^x (kf(u) - \theta - \beta v)v = 0, \tag{3.11}$$

$$\chi\phi'(u(x_0))u'(x_0)v(x_0) < k\int_0^{x_0} f(u)v = ku'(x_0).$$

Whence

$$v(x_0) < k/(\chi \min_{[0,1]} \phi'(u)). \tag{3.12}$$

To show (iii), assume $kf(1) - \theta - \beta v(1) < 0$. Since $kf(u(x)) - \theta - \beta v(x)$ must change sign on $[0,1]$, there exists $b \in (0,1)$ such that $kf(u(x)) - \theta - \beta v(x)$ is negative on $(b,1]$, equal to zero at $x = b$. Then on $(b,1)$, $(\lambda v' - \chi\phi'(u)u'v)' > 0$, and by the boundary condition at $x = 1$, $\lambda v' - \chi\phi'(u)u'v < 0$. So on $(b,1]$, $(v\exp(-\phi(u)\chi/\lambda))' < 0$ and hence

$$v(x) < e^{\frac{\chi}{\lambda}(\phi(u(x)) - \phi(u(b)))}v(b),\ x \in (b,1].$$

Since $v(b) = (kf(u(b)) - \theta)/\beta$, we see that $v(1)$ remains bounded.

Proof of (i) *of Theorem* 3.3. By (i) of Lemma 3.5 and the u-equation in (3.1), $\{u_i\}$ and $\{u_i'\}$ are equicontinuous on $[0,1]$. So after passing to a subsequence $\chi_n \to \infty$,

$$u_i \to \text{some } u_\infty \text{ in } C^1([0,1]). \tag{3.13}$$

There are two cases to consider.

Case 1. Each v_i is monotone increasing on $[0,1]$. Then as in the proof of Theorem 2.1, by Helly's theorem, after passing to another subsequence, we have $v_i \to v_\infty$ pointwise on $[0,1)$. Similarly as in the proof of (ii) of Theorem 2.1, we have $u_\infty'(x)v_\infty(x) = 0$ for $x \in [0,1)$. Combining this with

$$u_\infty'(x) = \int_0^x f(u_\infty(s))v_\infty(s)ds,\ x \in [0,1), \tag{3.14}$$

we have $u_\infty' \equiv 0$ and hence $u_\infty \equiv 1$, $v_\infty \equiv 0$ on $[0,1)$.

Case 2. There exists a subsequence of v_i (still denoted by itself) such that each v_i is not monotone increasing (not monotone decreasing either because of the boundary condition at $x = 1$). Let y_i be the largest local maximum point of v_i in $[0,1)$ and let z_i be the next local minimum point.

Subcase 1. $y_i \to 1$ as $i \to \infty$. In this case, (3.12) implies that $v_i \to 0$ in $C^0_{\text{loc}}([0,1))$. Then from (3.14) we have $u_\infty' \equiv 0$ and hence $u_\infty \equiv 1$ on $[0,1]$.

Subcase 2. $y_i \to$ some $y_\infty \in [0,1)$, $z_i \to z_\infty \in [y_\infty, 1)$ as $i \to \infty$. Again, (3.12) implies $v_i \to 0$ in $C^0_{\text{loc}}([0,y_\infty))$. On the other hand, since v_i is monotone increasing on $[z_i,1]$, we can use the arguments in Case 1 to show $v_i \to 0$ in $C^0_{\text{loc}}((z_\infty,1))$. Now it is clear that $v_i \to 0$ in $C^0_{\text{loc}}([0,1))$ and $u_\infty \equiv 1$.

We have shown that $u_i \to 1$ in $C^1([0,1])$ and $v_i \to 0$ in $C^0_{\text{loc}}([0,1))$ (without passing to a subsequence). This and (i) of Lemma 3.5 imply $\int_0^1 v_i(x)dx \to 0$. Moreover, by (3.2), it is easy to see that for large i the maximum of v_i is no less than

$(kf(1)-\theta)/2\beta$. Then in view of (3.12), the maximum must be achieved at $x=1$ for large i. Therefore we also have $v_i(1) \not\to 0$.

Proof of (ii) *and* (iii) *of Theorem* 3.3. The proof is similar to that for the corresponding parts of Theorem 2.1 ((iii) and (iv)). We just need to note that the boundedness of v needed for our present situation is guaranteed by (ii) and (iii) of Lemma 3.5.

Proof of (iv) *of Theorem* 3.3. As in the proof of part (i), after passing to a subsequence, $u_i \to$ some u_0 in $C^1([0,1])$. We want to show that along this sequence,

$$v_i \to v_0 \equiv (kf(u_0)-\theta)/\beta \text{ pointwise on } [0,1). \tag{3.15}$$

This will be done by showing that for any $x_0 \in [0,1)$, after passing to yet another subsequence, $v_i(x_0) \to v_0(x_0)$. To this end, define $\bar{u}_i(x) = u_i(\sqrt{\lambda_i}x + x_0)$, $\bar{v}_i(x) = v_i(\sqrt{\lambda_i}x + x_0)$, $x \in ([0,1]-x_0)/\sqrt{\lambda_i}$. We assume $x_0 > 0$—the case $x_0 = 0$ can be handled by first making the even extension of u_i and v_i. $\bar{v}_i$ satisfies on $([0,1]-x_0)/\sqrt{\lambda_i}$,

$$\begin{aligned} &\bar{v}_i'' - \frac{\chi_i}{\sqrt{\lambda_i}}\phi'(\bar{u}_i)u_i'(\sqrt{\lambda_i}x+x_0)\bar{v}_i' - \chi_i\phi''(\bar{u}_i)\bar{v}_i(u_i')^2(\sqrt{\lambda_i}x+x_0) \\ &\quad - \chi_i\phi'(\bar{u}_i)f(\bar{u}_i)\bar{v}_i^2 + (kf(\bar{u}_i)-\theta-\beta\bar{v}_i)\bar{v}_i = 0. \end{aligned} \tag{3.16}$$

Claim 1. For any $a \in (0,1)$, v_i remains bounded on $[0,a]$ as $\lambda_i \to 0$. To see this, we observe that by (ii) of Lemma 3.5, the claim is true in any of the following cases: (i) the maximum of v_i on $[0,a]$ is achieved in $[0,a)$; (ii) the maximum is achieved at $x=a$ which is also a local maximum of v_i; (iii) v_i has a local maximum in $(a,1)$ with the function value greater than $v_i(a)$. Thus the only case left to consider is the one in which v_i is increasing on $[a,1]$ and $v_i(a)$ is unbounded. In this case the L^2-norm of v_i is unbounded, contradicting (i) of Lemma 3.5. Claim 1 is proved.

By Claim 1, $\bar{v}_i$ remains bounded on an interval of the form $(-\delta,\delta)/\sqrt{\lambda_i}$, which expands to the whole real line as $\lambda_i \to 0$. Applying the interior L^2-estimates of elliptic equations to (3.16), we have that for any fixed $M > 0$, $\|\bar{v}_i\|_{H^2(-M,M)}$ is bounded. Since $H^2(-M,M) \hookrightarrow C^1([-M,M])$, we see from (3.16) that $\|\bar{v}_i\|_{C^2([-M,M])}$ is bounded. Now differentiating (3.16), we have that $\|\bar{v}_i\|_{C^3([-M,M])}$ is also bounded. By a diagonalization argument and the Azela–Ascoli theorem, we obtain that after passing to a subsequence, $\bar{v}_i \to$ some $\bar{v}_0$ in $C^2_{\text{loc}}(\mathbb{R})$ as $\lambda_i \to 0$. Since $\chi_i/\sqrt{\lambda_i}$ is bounded, we can extract another sequence $\lambda_i \to 0$ such that $\chi_i/\sqrt{\lambda_i} \to a \geq 0$. Now sending $\lambda_i \to 0$ in (3.16), we are led to

$$\bar{v}_0'' - a\phi'(u_0(x_0))u_0'(x_0)\bar{v}_0' + (kf(u_0(x_0))-\theta-\beta\bar{v}_0)\bar{v}_0 = 0,\ x \in (-\infty,\infty), \tag{3.17}$$

where $\bar{v}_0 \geq 0$ is bounded on $(-\infty,\infty)$ because of Claim 1. We want to show $\bar{v}_0 \equiv (kf(u_0(x_0))-\theta)/\beta$.

Claim 2. $u_0(x) > c$ for $x \in [0,1]$, where $kf(c) = \theta$. We first prove $u_0 \geq c$. Suppose this is not true. Then there exists $x_0 \in (0,1]$ such that $u_0(x) < c$ for $0 \leq x < x_0$ and $u_0(x_0) = c$ or $x_0 = 1$. Since v_i is bounded in $L^2([0,1])$, after passing to a subsequence, $v_i \to$ some $\widetilde{v}_0$ weakly in $L^2([0,1])$ as $i \to \infty$. Obviously $\widetilde{v}_0 \geq 0$. For any $\varphi \in C_0^\infty((0,1))$, we have

$$\lambda_i \int_0^1 v_i\varphi'' + \chi_i \int_0^1 \phi'(u_i)u_i'v_i\varphi' + \int_0^1 (kf(u_i)-\theta-\beta v_i)v_i\varphi = 0.$$

Sending $\lambda_i \to 0$, we have $\lim_{i\to\infty}\int_0^1 (kf(u_i)-\theta-\beta v_i)v_i\varphi = 0$. Take $\varphi \geq 0$. Define space $L^2_\varphi(0,1)$ with the inner product $(g,h) = \int_0^1 gh\varphi dx$. Then in $L^2_\varphi(0,1)$, $v_i \to \widetilde{v}_0$

weakly and hence $\|\widetilde{v}_0\|^2_{L^2_\varphi(0,1)} \leq \liminf_{i\to\infty} \|v_i\|^2_{L^2_\varphi(0,1)}$. Now

$$\begin{aligned}\int_0^1 (kf(u_0)-\theta)\widetilde{v}_0\varphi &= \lim_{i\to\infty}\int_0^1 (kf(u_i)-\theta)v_i\varphi, \\ &= \beta\lim_{i\to\infty}\int_0^1 v_i^2\varphi \geq \beta\int_0^1 \widetilde{v}_0^2\varphi,\end{aligned}$$

i.e., $\int_0^1 (kf(u_0)-\theta-\beta\widetilde{v}_0)\widetilde{v}_0\varphi \geq 0$. Therefore,

$$(kf(u_0)-\theta-\beta\widetilde{v}_0)\widetilde{v}_0 \geq 0 \ \text{ a.e. on } \ (0,1). \tag{3.18}$$

Since $kf(u_0)-\theta<0$ on $[0,x_0)$, $\widetilde{v}_0 \equiv 0$ a.e. on $(0,x_0)$. Because

$$u_0'(x) = \int_0^x f(u_0(s))\widetilde{v}_0(s)ds, \ x\in[0,1], \tag{3.19}$$

$u_0' \equiv 0$ on $[0,x_0]$. Thus $u_0 \equiv \text{const} < c$ on $[0,x_0]$ and hence $x_0 = 1$. This contradicts $u_0'(1) = h(1-u_0(1))$.

We have shown $u_0 \geq c$ on $[0,1]$ and combining this with (3.18), we have $\widetilde{v}_0 \leq (kf(u_0)-\theta)/\beta$ a.e. on $(0,1)$. Now (3.19) yields

$$u_0'(x) \leq \int_0^x f(u_0(x))(kf(u_0)-\theta)/\beta\, dx, \ x\in[0,1].$$

Define $w(x) = u_0(x) - c$. Then $w \geq 0$, is nondecreasing on $[0,1]$, and satisfies for some positive function $c(x)$ the following:

$$w'(x) \leq \int_0^x c(s)w(s)ds \leq w(x)\int_0^x c(s)ds.$$

It follows from this that if $w(0) = 0$, then $w \equiv 0$ on $[0,1]$, which would contradict $u_0'(1) = h(1-u_0(1))$. Thus $w(x) \geq w(0) > 0$ for $x \in [0,1]$. Claim 2 is proved.

Claim 3. $\bar{v}_0(x) > 0$ for $x \in \mathbb{R}$. If $\bar{v}_0 = 0$ somewhere, then by the strong maximum principle $\bar{v}_0 \equiv 0$. Then the minimum of v_i on $[0,1]$ shrinks to zero as $i \to \infty$. At a minimum point $\bar{x}_0$ of v_i, $v_i' = 0$ and $v_i'' \geq 0$. Then by (3.10), we have that at $\bar{x}_0$,

$$kf(u_i) - \theta - \beta v_i - \chi_i\phi''(u_i)(u_i')^2 - \chi_i\phi'(u_i)f(u_i)v_i \leq 0,$$

i.e.,

$$\frac{kf(u_i) - \chi_i\phi''(u_i)(u_i')^2 - \theta}{\beta + \chi_i\phi'(u_i)f(u_i)} \leq v_i.$$

Thus $kf(u_0(0)) - \theta \leq 0$, contradicting Claim 2. Claim 3 is proved.

Claim 4. $I(x) \equiv kf(u_0(x_0)) - \theta - \beta\bar{v}_0(x) \geq 0$ on $\mathbb{R}$. Suppose $I(x) < 0$ on an interval (r,s). Then (3.17) implies (writing $b = a\phi'(u_0(x_0))u_0'(x_0)$),

$$\bar{v}_0'' - b\bar{v}_0' > 0, \ \text{ i.e., } \ (e^{-bx}\bar{v}_0')' > 0 \ \text{ on } \ (r,s).$$

If there exists $x_0 \in (r,s)$ such that $\bar{v}_0'(x_0) \geq 0$, then $\bar{v}_0' > 0$ on (x_0,s). This implies that $I < 0$ on (x_0,∞) which in return implies $\bar{v}_0' > 0$ on (x_0,∞). Now we see that $\bar{v}_0'' > \text{const} > 0$ on (x_0,∞) which contradicts the boundedness of $\bar{v}_0$. If $\bar{v}_0'(x_0) < 0$, we would be led to a similar contradiction again. Claim 4 is proved.

Claim 5. $I(x) \leq 0$ on $\mathbb{R}$. Suppose $I > 0$ on an interval (r, s). Then $(e^{-bx}\bar{v}_0')' < 0$ on (r, s). If there exists $x_0 \in (r, s)$ such that $\bar{v}_0'(x_0) \geq 0$, then $\bar{v}_0' > 0$ on (r, x_0). Hence $I > 0$, $\bar{v}_0' > 0$ on $(-\infty, x_0)$. Then $\lim_{x\to-\infty} \bar{v}_0(x) = 0$, which implies that $\min_{[0,1]} v_i \to 0$ as $i \to \infty$. As in Claim 3, this is impossible. If $\bar{v}_0'(x_0) < 0$, we will be led to a similar contradiction again.

We have shown $\bar{v}_0 \equiv (kf(u_0(x_0)) - \theta)/\beta$ and hence (3.15). By the equation $u_i'(x) = \int_0^x f(u_i(s))v_i(s)ds$, Lebesgue's dominated theorem, and Claim 1, we have

$$u_0'(x) = \int_0^x f(u_0(s))(kf(u_0(s)) - \theta)/\beta\, ds, \ x \in [0, 1).$$

Therefore u_0 satisfies (3.9). By the comparison principle, (3.9) has at most one solution satisfying $kf(u_0) - \theta \geq 0$, i.e., $u_0 \geq c$ on $[0, 1]$. Thus, without passing to a subsequence, $u_i \to u_0$ in $C^1([0, 1])$ and $v_i \to v_0$ pointwise on $[0, 1)$ as $i \to \infty$. The proof of (iv) is complete.

Proof of (v) *of Theorem* 3.3. We first show that in the current case $\phi(u) = u$, v is strictly increasing on $[0, 1]$. Otherwise, there exist a and b in $[0, 1)$ with $a < b$ such that $v(a) \geq v(b)$ and $v(a)$ and $v(b)$ are local maximum and local minimum, respectively. Then by inspecting (3.10) ($\phi'' = 0$), we have

$$kf(u(a)) - \theta - \beta v(a) - \chi f(u(a))v(a) \geq 0 \geq kf(u(b)) - \theta - \beta v(b) - \chi f(u(b))v(b).$$

Thus

$$\frac{kf(u(a)) - \theta}{\beta + \chi f(u(a))} \geq v(a) \geq v(b) \geq \frac{kf(u(b)) - \theta}{\beta + \chi f(u(b))}.$$

This is impossible because $u(a) < u(b)$ and hence $f(u(a)) < f(u(b))$. We have shown that v is strictly increasing on $[0, 1]$. Now as before, after passing to a subsequence, $v_i \to v_0$ pointwise on $[0, 1)$, $u_i \to u_0$ in $C^1([0, 1])$ as $i \to \infty$.

We first consider the case $\chi_i \to 0$ as $\lambda_i \to 0$. As in the proof of (i) of Theorem 2.1, we have $(kf(u_0) - \theta - \beta v_0)v_0 = 0$ a.e. in $(0, 1)$. By exactly the same arguments in Claim 2 of the proof of (iv), $kf(u_0(x)) - \theta > 0$, $x \in [0, 1]$. This and the arguments in Claim 3 imply $v_0 > 0$. Thus $kf(u_0) - \theta - \beta v_0 = 0$ a.e. on $(0, 1)$. Since both u_0 and v_0 are nondecreasing and u_0 is continuous, the above equation must hold everywhere in $[0, 1)$. In particular, v_0 is continuous on $[0, 1)$. Now it is easy to prove that the convergence $v_i \to v_0$ is uniform outside any fixed neighborhood of $x = 1$. It is also easy to see that u_0 satisfies (3.9) (with $kf(u_0) - \theta = \beta v_0 > 0$). Since u_0 is unique, we have $u_i \to u_0$ in $C^1([0, 1])$, $v_i \to v_0 \equiv (kf(u_0) - \theta)/\beta$ uniformly outside any neighborhood of $x = 1$, without passing to a subsequence.

Now we consider the case when $\chi_i \to \chi_0 > 0$ as $\lambda_i \to 0$. By using the arguments that lead to (2.6), we have

$$\chi_0 u_0'(x)v_0(x) = \int_0^x (kf(u_0(s)) - \theta - \beta v_0(s))v_0(s)ds, \text{ a.e. in } (0, 1). \tag{3.20}$$

From this, we see that $u_0(x) > c$, i.e., $kf(u_0(x)) - \theta > 0$ for $x \in [0, 1)$. Then as can be seen as follows, $v_0(0) > 0$ and hence $v_0 > 0$ on $[0, 1)$. Suppose $v_0(0) = 0$. Since $v_i''(0) \geq 0$ and $v_i'(0) = 0$, by (3.10) we have $kf(u_i(0)) - \theta - \beta v_i(0) - \chi_i f(u_i(0))v_i(0) \leq 0$. Sending $i \to \infty$, we obtain $kf(u_0(0)) - \theta \leq 0$. We have a contradiction. Now $u_0'(x) = \int_0^x f(u_0(s))v_0(s)ds > 0$ for $x \in (0, 1]$. Then (3.20) implies that v_0 is continuous on $(0, 1)$. If we define v_0 at $x = 0, 1$ by the one-sided limits, v_0 is actually continuous on

$[0,1]$ with $v_0 > 0$. Then u_0'' exists on $[0,1]$ and $u_0''(x) = f(u_0(x))v_0(x) > 0$ for $x \in [0,1]$. Now it follows from (3.20) that $(kf(u_0(x)) - \theta - \beta v_0(x))v_0(x) > 0$, $x \in [0,1]$. Thus $kf(u_0) > \theta + \beta v_0$ on $[0,1]$.

Since v_0 is continuous, the convergence $v_i \to v_0$ is uniform outside any neighborhood of $x = 1$. At last, we show $v_i(1) \to \infty$ (this does not contradict $v_0(1) < \infty$). For any $\delta \in (0,1)$, we write

$$\beta \int_0^\delta v_i^2(x)dx = \int_0^1 (kf(u_i(x)) - \theta)v_i(x)dx - \beta \int_\delta^1 v_i^2(x)dx.$$

Sending $i \to \infty$, by Lebesgue's dominated convergence theorem, we have

$$\beta \int_0^\delta v_0^2(x)dx = \int_0^1 (kf(u_0(x)) - \theta)v_0(x)dx - \beta \lim_{i\to\infty} \int_\delta^1 v_i^2(x)dx.$$

Now

$$\begin{aligned}\beta \lim_{\delta\to 1}\lim_{i\to\infty} \int_\delta^1 v_i^2(x)dx &= \int_0^1 (kf(u_0(x)) - \theta)v_0(x)dx - \beta \int_0^1 v_0^2(x)dx \\ &= \int_0^1 (kf(u_0(x)) - \theta - \beta v_0(x))v_0(x)dx > 0.\end{aligned}$$

Thus $v_i(1) \to \infty$ as $i \to \infty$. The proof of Theorem 3.3 is complete.

4. Time-dependent solutions. In this section, we study the global existence and boundedness of solutions of (1.5) ($\beta = 0$ allowed and hence the Malthusian case included). Some estimates proved in the process will be useful in the study of the stability of the trivial steady state $(1,0)$ in the next section.

The local existence and uniqueness of the solution of (1.5) with the initial condition

$$u(x,0) = u_0(x),\ v(x,0) = v_0(x) \tag{4.1}$$

follow from [Am 1, Am 2].

THEOREM 4.1 (local existence). *Extend f and ϕ so that*

$$f \in C^3(\mathbb{R}) \quad \textit{and} \quad \phi \in C^5(\mathbb{R}).$$

(i) *For any u_0, $v_0 \in H^1(0,1)$, there exists a unique maximal solution $(u(x,t), v(x,t))$ defined on $[0,1] \times [0, T_{(u_0,v_0)})$ with $0 < T_{(u_0,v_0)} \le \infty$ such that*

$$\begin{gathered}(u(\cdot,t), v(\cdot,t)) \in C([0,T_{(u_0,v_0)}),\ H^1(0,1) \times H^1(0,1)),\\ (u,v) \in C^{2+2\epsilon,1+\epsilon}_{\text{loc}}([0,1] \times (0, T_{(u_0,v_0)}))\end{gathered}$$

for any $0 < \epsilon < \frac{1}{4}$.

(ii) *Let $\varphi(t,(u_0,v_0))$ be the unique solution described above. Then φ is a $C^{0,1}$-map from $\{(t,(u,v)) \mid (u,v) \in H^1(0,1) \times H^1(0,1),\ 0 < t < T_{(u,v)}\}$ to $H^1(0,1) \times H^1(0,1)$.*

(iii) *If $\|(u,v)(\cdot,t)\|_{L^\infty(0,1)}$ is bounded for $t \in [\delta, T_{(u_0,v_0)})$, δ small, then $T_{(u_0,v_0)} = \infty$, i.e., (u,v) is global in time. Furthermore, $(u,v) \in C^\rho([\delta,\infty),\ C^{2(1-\sigma)}[0,1] \times C^{2(1-\sigma)}[0,1])$ for any $0 \le \rho \le \sigma \le 1$.*

Proof. Let $w = 1 - u$. Then (1.5) is written as follows:

$$\begin{cases} \binom{w}{v}_t = \Big[A(w,v)\binom{w}{v}_x\Big]_x + \binom{f(1-w)v}{(kf(1-w)-\theta-\beta v)v}, \\ A(w,v)\binom{w}{v}_x + H(x)\binom{w}{v} = 0 \ \text{ at } \ x = 0, 1, \end{cases} \tag{4.2}$$

where

$$A(w,v) = \begin{pmatrix} 1 & 0 \\ \chi\phi'(1-w)v & \lambda \end{pmatrix}, \ H(x) = \begin{pmatrix} hx & 0 \\ 0 & 0 \end{pmatrix}.$$

Since the eigenvalues of A are positive, (4.2) is "normally parabolic" [Am 1]. Then (i) and (ii) follow from [Am 1, Theorems 7.3 and 9.3]. (iii) follows from [Am 2, Theorem 5.2] because (4.2) is a "triangular system."

PROPOSITION 4.2 (positivity). *Suppose in Theorem* 4.1, $1 \ge u_0 \ge 0$, $v_0 > 0$ *on* $[0,1]$. *Then* $1 > u > 0$ *and* $v > 0$ *on* $[0,1] \times [0,T)$, $T = T_{(u_0,v_0)}$.

REMARK 4.3. Later we shall handle the case $v_0 \ge 0$ (see the discussion preceding Theorem 4.8).

Proof. Since $(u(\cdot,t), v(\cdot,t)) \in C([0,T), H^1(0,1) \times H^1(0,1)) \subset C([0,T), C[0,1] \times C[0,1])$, if $v > 0$ on $[0,1] \times [0,T)$ is untrue, then there exists $(x_0,t_0) \in [0,1] \times (0,T)$ such that $v(x_0,t_0) = 0$, $v(x,t) > 0$ for $(x,t) \in [0,1] \times [0,t_0)$. v obviously satisfies

$$\begin{cases} v_t = \lambda v_{xx} + b(x,t)v_x + c(x,t)v, \ (x,t) \in [0,1] \times (0,T), \\ \lambda v_x = d(x,t)v \ \text{ at } \ x = 0, 1, \end{cases} \tag{4.3}$$

where b, c, and $d(x,t)$ are continuous on $[0,1] \times (0,T)$. If $x_0 \in (0,1)$, then by the strong maximum principle, $v \equiv 0$ on $[0,1] \times [0,t_0]$, which is impossible. On the other hand, if x_0 is either 0 or 1, then by the boundary condition in (4.3), $v_x(x_0,t_0) = 0$. This also is impossible by the Hopf boundary point lemma. We have thus shown $v > 0$ on $[0,1] \times [0,T)$. Now observe that $\underline{u} \equiv 0$ and $\bar{u} \equiv 1$ are strict lower and upper solutions of the u-equation with the boundary condition in (1.5), because $v > 0$. By the comparison principle, $0 < u < 1$ on $[0,1] \times (0,T)$. Proposition 4.2 is proved.

We now proceed to establish the L^∞-bound of v under the condition on the initial value (u_0,v_0) as in Proposition 4.2. Then in light of (iii) of Theorem 4.1, the global existence of the solution of (1.5) and (4.1) follows. The L^∞-bound of v will be obtained through series of lemmas.

LEMMA 4.4. *Suppose* (u_0,v_0) *satisfies the condition in Theorem* 4.1 *and Proposition* 4.2. *Let* (u,v) *be the unique (positive) solution of* (1.5) *and* (4.1). *Then*

$$\int_0^1 v(x,t)dx \le \int_0^1 (v_0(x) + ku_0(x))dx + k(h+\theta)/\theta, \ 0 \le t < T, \tag{4.4}$$

where $T = T_{(u_0,v_0)}$.

Proof. Let $\bar{u}(t) = \int_0^1 u(x,t)dx$, $\bar{v}(t) = \int_0^1 v(x,t)dx$. Integrating (1.5) on $[0,1]$, we have that for $0 < t < T$,

$$\begin{aligned} \bar{u}'(t) &= -\int_0^1 f(u)v\,dx + h(1 - u(1,t)) \le -\int_0^1 f(u)v\,dx + h, \\ \bar{v}'(t) &\le k\int_0^1 f(u)v\,dx - \theta\bar{v}(t) \le kh - k\bar{u}'(t) - \theta\bar{v}(t), \\ &\le -k\bar{u}'(t) - \theta(\bar{v}(t) + k\bar{u}(t)) + kh + \theta k. \end{aligned}$$

Thus

$$\bar{v}(t) + k\bar{u}(t) \leq e^{-\theta t}(\bar{v}(0) + k\bar{u}(0)) + (kh + \theta k)(1 - e^{-\theta t})/\theta.$$

This completes the proof of Lemma 4.4.

LEMMA 4.5. *Under the condition on (u_0, v_0) in Lemma 4.4 for any small $\delta > 0$, we have*

$$\int_0^1 v^2(x,t)dx \leq 2\int_0^1 v^2(x,\delta)dx + \mu, \ \ \delta \leq t < T,$$

where $T = T(u_0, v_0)$ and μ is a positive constant depending on all the constants in (1.5), *f, ϕ, v_0, and $\|u(x,\delta)\|_{H^2(0,1)}$.*

Proof. We shall use "energy estimates" to get the L^2-bound for v. But because of the chemotactic flux term in the v-equation, we first need to estimate the L^2-norm of u_x.

For $2 > p > 1$, let $X = L^p(0,1)$, $\Delta = \frac{d^2}{dx^2}$ with domain $D(\Delta) = W^{2,p}(0,1) \cap \{u'(0) = 0 = u'(1) + hu(1)\}$. Then Δ generates a linear analytic semigroup $e^{t\Delta}$ on the Banach space X. Since the spectrum of Δ lies on the negative real line, away from 0, $\|e^{t\Delta}\| \leq ce^{-at}$ for $t \geq 0$ and for some positive constant a [H, Thm. 1.3.4]. Also, the fractional space $X^\alpha \hookrightarrow H^1(0,1)$ if $\alpha > (\frac{1}{2} + \frac{1}{p})/2$ [H, Thm. 1.6.1]. We take p close to 1 so α can be taken close to 3/4. By [H, Thm. 1.4.3], $\|e^{t\Delta}\|_\alpha \leq C_\alpha t^{-\alpha}e^{-at}$. Let $w = 1 - u$; then w is the classical solution of

$$\begin{cases} w_t = w_{xx} + f(u)v, \ \ x \in [0,1], \ \ \delta \leq t < T, \\ w_x(0,t) = 0 = w_x(1,t) + hw(1,t), \\ w(x,\delta) = 1 - u(x,\delta). \end{cases} \tag{4.5}$$

Therefore w satisfies the following variation of constants formula:

$$w(\cdot,t) = e^{(t-\delta)\Delta}(1 - u(\cdot,\delta)) + \int_\delta^t e^{(t-s)\Delta}f(u(\cdot,s))v(\cdot,s)ds, \ \ \delta \leq t < T.$$

Now for $\delta \leq t < T$,

$$\begin{aligned} \|u_x(\cdot,t)\|_{L^2(0,1)} &\leq \|w(\cdot,t)\|_{H^1(0,1)} \leq \|w(\cdot,t)\|_{X^\alpha} \\ &\leq \|e^{(t-\delta)\Delta}(1 - u(\cdot,\delta))\|_{X^\alpha} + \int_\delta^t \|e^{(t-s)\Delta}f(u(\cdot,s))v(\cdot,s)\|_{X^\alpha}ds \\ &\leq ce^{-a(t-\delta)}\|1 - u(\cdot,\delta)\|_{X^\alpha} \\ &\quad + \int_\delta^t c_\alpha(t-s)^{-\alpha}e^{-a(t-s)}\|f(u(\cdot,s))v(\cdot,s)\|_X ds. \end{aligned}$$

By the Hölder's inequality,

$$\|v(\cdot,s)\|_X = \|v(\cdot,s)\|_{L^p(0,1)} \leq (\bar{v}(s))^{\frac{2-p}{p}} \left(\int_0^1 v^2(x,s)dx\right)^{\frac{p-1}{p}}.$$

Let $K(t) = \max_{\delta \leq s \leq t} \int_0^1 v^2(x,s)dx$. The above estimates together with Lemma 4.4 imply that for $\delta \leq t < T$,

$$\|u_x(\cdot,t)\|_{L^2(0,1)} \leq c_1 + c_2(K(t))^{\frac{p-1}{p}}, \tag{4.6}$$

where c_1 depends on $\|u(\cdot,\delta)\|_{H^2(0,1)}$, and c_2 on f, v_0, k, h, and θ.

In the following, we use these inequalities:

$$a^r b^{1-r} \le ra + (1-r)b \quad \text{(Young's inequality)}, \tag{4.7}$$

$$\|v\|_{L^\infty(0,1)} \le c\left(\bar{v} + \bar{v}^{1/3}\left(\int_0^1 v_x^2 dx\right)^{1/3}\right), \tag{4.8}$$

$$\int_0^1 v^2 dx \le \varepsilon \int_0^1 v_x^2 dx + (c\varepsilon^{-1/2} + 1)\bar{v}^2 \tag{4.9}$$

(Garliardo–Ladyzenskaya–Nirenberg inequality).

(For (4.8) and (4.9), see [LSU, Thm. 2.2 and Remark 2.1].)

We now multiply v to the v-equation in (1.5) and integrate by parts on $[0,1]$. We obtain that for $\delta \le t < T$,

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_0^1 v^2(x,t)dx \\
&\quad \le -\int_0^1 (\lambda v_x - \chi\phi'(u)u_x v)v_x dx + \int_0^1 kf(u)v^2 dx \\
&\quad \le -\lambda\int_0^1 v_x^2 dx + \chi\int_0^1 (\phi'(u)v)u_x v_x dx + kf(1)\int_0^1 v^2 dx \\
&\quad \overset{(4.8)}{\le} -\lambda\int_0^1 v_x^2 dx + c_3\left(\bar{v} + \bar{v}^{1/3}\left(\int_0^1 v_x^2\right)^{1/3}\right)\left(\int_0^1 u_x^2 dx\right)^{1/2}\left(\int_0^1 v_x^2 dx\right)^{1/2} \\
&\qquad\quad + kf(1)\int_0^1 v^2 dx, \\
&\quad \overset{(4.7)}{\le} -\lambda\int_0^1 v_x^2 dx + \frac{\lambda}{4}\int_0^1 v_x^2 dx + \frac{1}{\lambda}c_3^2\bar{v}^2\int_0^1 u_x^2 dx \\
&\qquad\quad + c_3\bar{v}^{1/3}\left(\int_0^1 u_x^2 dx\right)^{1/2}\left(\int_0^1 v_x^2 dx\right)^{5/6} + kf(1)\int_0^1 v^2 dx, \\
&\quad \overset{(4.7)}{\le} \frac{-3\lambda}{4}\int_0^1 v_x^2 dx + \frac{c_3^2\bar{v}^2}{\lambda}\int_0^1 u_x^2 dx + \frac{\lambda}{4}\int_0^1 v_x^2 dx \\
&\qquad\quad + c_4\bar{v}^2\left(\int_0^1 u_x^2 dx\right)^3 + kf(1)\int_0^1 v^2 dx, \\
&\quad \le -\frac{\lambda}{2}\int_0^1 v_x^2 dx + c_5\bar{v}^2 + c_5\bar{v}^2\left(\int_0^1 u_x^2 dx\right)^3 + kf(1)\int_0^1 v^2 dx, \\
&\quad \overset{(4.9)}{\le} -\frac{\lambda}{2\varepsilon}\left(\int_0^1 v^2 dx - (c\varepsilon^{-1/2} + 1)\bar{v}^2\right) + c_5\bar{v}^2 + c_5\bar{v}^2\left(\int_0^1 u_x^2 dx\right)^3 \\
&\qquad\quad + kf(1)\int_0^1 v^2 dx, \\
&\quad = \left(kf(1) - \frac{\lambda}{2\varepsilon}\right)\int_0^1 v^2 dx + c_6\bar{v}^2 + c_5\bar{v}^2\left(\int_0^1 u_x^2 dx\right)^3.
\end{aligned}$$

Taking $\varepsilon = \lambda/(3kf(1))$ and using (4.6), we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 v^2(x,t)dx &\le \frac{-kf(1)}{2}\int_0^1 v^2dx + c_6\bar{v}^2 + c_5\bar{v}^2(c_1 + c_2(K(t))^{(p-1)/p})^6,\\ &\le \frac{-kf(1)}{2}\int_0^1 v^2dx + c_7\bar{v}^2 + c_8\bar{v}^2(K(t))^{6(p-1)/p},\end{aligned}$$

for which it follows that for $\delta \le t < T$,

$$\int_0^1 v^2(x,t)dx \le \int_0^1 v^2(x,\delta)dx + c_9(1 + (K(t))^{6(p-1)/p}).$$

Thus

$$K(t) \le \int_0^1 v^2(x,\delta)dx + c_9(1 + (K(t))^{6(p-1)/p}).$$

Now we take $p > 1$ but close to 1 such that $6(p-1)/p < 1$. Lemma 4.5 follows.

Lemma 4.6. *Let the initial value (u_0, v_0) be as in Lemma 4.4. Then*

$$\|u_x(\cdot,t)\|_{L^\infty(0,1)} \le L,\ \ \delta \le t < T,$$

where the constant L depends on the items that μ in Lemma 4.5 also depends on, plus the L^2-norm of $v(x,\delta)$.

Proof. Take $p = 2$ at the beginning of the proof of Lemma 4.5. Then $X^\alpha \hookrightarrow C^\nu[0,1]$, where ν is any number in $(0, 2\alpha - \frac{1}{2})$. Take $\alpha = 7/8$ so ν can be taken to be 1. Since $w = 1 - u$ satisfies the variation of constants formula below (4.5),

$$\begin{aligned}\|u'\|_{C[0,1]} &\le \|w(\cdot,t)\|_{X^\alpha}\\ &\le ce^{-a(t-\delta)}\|1 - u(\cdot,\delta)\|_{X^\alpha} + \int_\delta^t c_\alpha(t-s)^{-\alpha}e^{-a(t-s)}\|f(u(\cdot,s))v(\cdot,s)\|_X ds,\\ &\overset{\text{(Lemma 4.5)}}{\le} ce^{-a(t-\delta)}\|1 - u(\cdot,\delta)\|_{X^\alpha}\\ &\qquad + \left(2\int_0^1 v^2(x,\delta)dx + \mu\right) f(1)c_\alpha \int_\delta^t (t-s)^{-\alpha}e^{-a(t-s)}ds,\end{aligned}$$

for $\delta \le t < T$. This completes the proof of Lemma 4.6.

Lemma 4.7. *Let the initial value (u_0, v_0) be as in Lemma 4.4. Then $\|v(\cdot,t)\|_{L^\infty} \le c$, $\delta \le t < T$, where constant c depends on all items that L in Lemma 4.6 also depends on, plus $\|v(\cdot,\delta)\|_{L^\infty(0,1)}$.*

Proof. We use the Moser–Alikakos iteration [A]. For $s \ge 2$, multiplying the v-equation in (1.5) by v^{s-1} and integrating on $[0,1]$ by parts, we have

$$\begin{aligned}\frac{1}{s}\frac{d}{dt}\int_0^1 v^s(x,t)dx &\le -\lambda(s-1)\int_0^1 v_x^2 v^{s-2}dx + (s-1)\chi\int_0^1 \phi'(u)u_x v^{s-1}v_x dx\\ &\qquad + kf(1)\int_0^1 v^s dx,\\ &\overset{\text{(Lemma 4.6)}}{\le} \frac{-4\lambda(s-1)}{s^2}\int_0^1 (v^{s/2})_x^2 dx + c_1\int_0^1 v^{s/2}|(v^{s/2})_x|dx\\ &\qquad + kf(1)\int_0^1 v^s dx.\end{aligned}$$

Thus

$$\begin{aligned}
\frac{d}{dt}\int_0^1 v^s(x,t)dx &\le \frac{-4\lambda(s-1)}{s}\int_0^1 (v^{s/2})_x^2 dx + c_1 s\left(\int_0^1 v^s dx\right)^{1/2}\left(\int_0^1 (v^{s/2})_x^2 dx\right)^{1/2}\\
&\qquad + kf(1)s\int_0^1 v^s dx,\\
&\le -2\lambda\int_0^1 (v^{s/2})_x^2 dx + \lambda\int_0^1 (v^{s/2})_x^2 dx + \frac{c_1^2 s^2}{4\lambda}\int_0^1 v^s dx\\
&\qquad + kf(1)s\int_0^1 v^s dx,\\
&\le -\lambda\int_0^1 (v^{s/2})_x^2 dx + c_2 s^2\int_0^1 v^s dx,\\
&\overset{(4.9)}{\le} -\lambda\left(\frac{1}{\varepsilon}\int_0^1 v^s dx - \frac{(c_3\varepsilon^{-1/2}+1)}{\varepsilon}\left(\int_0^1 v^{s/2}dx\right)^2\right) + c_2 s^2\int_0^1 v^s dx,\\
&\le -c_2 s^2\int_0^1 v^s dx + c_4 s^3\left(\int_0^1 v^{s/2}dx\right)^2 \quad \left(\text{taking } \varepsilon = \frac{\lambda}{2c_2 s^2}\right).
\end{aligned}$$

From this, we have that for $\delta \le t < T$,

$$\frac{d}{dt}\left(e^{c_2 s^2 t}\int_0^1 v^s(x,t)dx\right) \le c_4 s^3 e^{c_2 s^2 t}\sup_{\delta\le t<T}\left(\int_0^1 v^{s/2}(x,t)dx\right)^2,$$

$$\int_0^1 v^s(x,t)dx \le \|v(\cdot,\delta)\|_{L^\infty(0,1)}^s + c_5 s\sup_{\delta\le t<T}\left(\int_0^1 v^{s/2}(x,t)dx\right)^2. \tag{4.10}$$

Let $M(s) = \max(\|v(\cdot,\delta)\|_{L^\infty(0,1)},\ \sup_{\delta\le t<T}(\int_0^1 v^s(x,t)dx)^{1/s})$. Then (4.10) implies

$$M(s) \le (c_6 s)^{1/s} M(s/2) \quad\text{for}\quad s \ge 2. \tag{4.11}$$

Taking $s = 2^k,\ k = 1, 2, \ldots$, we obtain

$$\begin{aligned}
M(2^k) &\le c_6^{\frac{1}{2^k}} 2^{\frac{k}{2^k}} M(2^{k-1}),\\
&\le c_6^{\frac{1}{2^k}+\cdots+1} 2^{\frac{k}{2^k}+\cdots+\frac{1}{2}} M(1).
\end{aligned}$$

Sending $k \to \infty$, we have that for $\delta \le t < T$,

$$\begin{aligned}
\|v(\cdot,t)\|_{L^\infty(0,1)} &\le \lim_{k\to\infty} M(2^k),\\
&\le c_7 M(1).
\end{aligned}$$

This completes the proof of Lemma 4.7.

Now, by (iii) of Theorem 4.1, the solution (u,v) of (1.5) and (4.1) is global in t, provided that (u_0, v_0) satisfies the conditions in Theorem 4.1 and Proposition 4.2, i.e., u_0, $v_0 \in H^1(0,1)$, $1 \ge u_0 \ge 0$, $v_0 > 0$ on $[0,1]$. We point out that we can actually allow $v_0 \ge 0, \not\equiv 0$ on $[0,1]$. To see this, we take a sequence v_{0n} in $H^1(0,1)$ with $v_{0n} > 0$ on $[0,1]$ and $v_{0n} \to v_0$ in $H^1(0,1)$. Let (u_n, v_n) be the global positive solutions of

(1.5) and (4.1) (with v_0 replaced by v_{0n}). Then by (ii) of Theorem 4.1, for $x \in [0,1]$, $0 \le t < T_{(u_0,v_0)}$, $(u_n(x,t), v_n(x,t)) \to (u(x,t), v(x,t))$. Thus $1 \ge u \ge 0$, $v \ge 0$. By the strong maximum principle and the Hopf boundary point lemma, $1 > u > 0$, $v > 0$ on $[0,1] \times (0, T_{(u_0,v_0)})$. Note that the condition $1 \ge u_0 \ge 0$, $v_0 > 0$ was required in Lemmas 4.4–4.7 only for the reason of ensuring $1 > u > 0$, $v > 0$. Thus in the present situation, Lemmas 4.4–4.7 still apply and hence $T(u_0, v_0) = \infty$.

We have thus shown the main result of this section.

Theorem 4.8 (global existence and boundedness). *For any $u_0, v_0 \in H^1(0,1)$ satisfying $1 \ge u_0 \ge 0$, $v_0 \ge 0$, $\not\equiv 0$ on $[0,1]$, (1.5) and (4.1) have a unique positive global solution (u,v) such that*

(i) $(u(\cdot,t), v(\cdot,t)) \in C([0,\infty),\ H^1(0,1) \times H^1(0,1))$, $(u,v) \in C^{2+2\varepsilon,1+\varepsilon}_{\text{loc}}([0,1] \times [0,\infty))$;

(ii) $0 < u < 1$, $v > 0$ *and is bounded on* $[0,1] \times [0,\infty)$.

5. Stability and instability of steady states. The first result of this section deals with the stability of the trivial steady state $(u,v) = (1,0)$ of (1.5) (again $\beta = 0$ is allowed so that both Malthusian and logisitc cases are included).

Theorem 5.1. (i) *Suppose $kf(1) \le \theta$. Then in L^∞-topology $(u,v) = (1,0)$ attracts every positive solution of* (1.5) *whose initial value $(u(x,0), v(x,0)) = (u_0(x), v_0(x))$ satisfies the condition in Theorem* 4.8. *Furthermore, if $kf(1) < \theta$, then $\|v(\cdot,t)\|_{L^\infty} \le C\ \exp((kf(1)-\theta)t)$, $\|1-u(\cdot,t)\|_{L^\infty} \le C\ \exp(-\min(a, \theta - kf(1))t)$, $t \ge 0$, where a is any number less than the first eigenvalue α of $-d^2/dx^2$ with the boundary condition $u'(0) = 0 = u'(1) + hu(1)$; if $kf(1) = \theta$ and $\beta > 0$, then $\|v(\cdot,t)\|_{L^\infty} \le \frac{C}{\beta t+1}$, $\|1-u(\cdot,t)\|_{L^\infty} \le \frac{C}{\beta t+1}$, $t \ge 0$.*

(ii) *Suppose $kf(1) > \theta$. Then $(u,v) = (1,0)$ is unstable in the L^∞-topology.*

Proof of (i). Consider first the case $kf(1) < \theta$. Let $z = e^{(\theta - kf(1))t} v$. Then

$$(5.1)\qquad \begin{cases} z_t \le (\lambda z_x - \chi\phi'(u) u_x z)_x, \\ \lambda z_x - \chi\phi'(u) u_x z = 0 \ \text{ at } \ x = 0,1, \\ z(x,0) = v_0(x). \end{cases}$$

Integrating (5.1) on $[0,1]$, we see that

$$\int_0^1 z(x,t)dx \le \int_0^1 v_0(x)dx,\ t \ge 0.$$

This and the proof of Lemma 4.7 imply that $\|z(\cdot,t)\|_{L^\infty} \le \text{const } C$, i.e., $\|v(\cdot,t)\|_{L^\infty} \le C\ \exp((kf(1)-\theta)t)$, $t \ge 0$. We now use the comparison principle to obtain a decay rate for $\|1-u(\cdot,t)\|_{L^\infty}$. Observe $1-u$ is a subsolution of

$$(5.2)\qquad \begin{cases} w_t = w_{xx} + C\ \exp((kf(1)-\theta)t), \\ w_x(0,t) = 0 = w_x(1,t) + hw(1,t), \\ w(x,0) = 1 - u_0(x). \end{cases}$$

Define $\bar{w}(x,t) = K\psi(x)e^{-\min(a,\theta-kf(1))t}$, where $K > 0$ is a constant, ψ is a first eigenfunction of $-d^2/dx^2$ with the boundary condition $u'(0) = 0 = u'(1) + hu(1)$. Since $\psi(x) > 0$ on $[0,1]$, we see for a large enough K, $\bar{w}$ is a supersolution of (5.2). By the comparison principle, $0 \le 1 - u(x,t) \le \bar{w}(x,t)$, $x \in [0,1]$, $t \ge 0$. Hence $\|1-u(\cdot,t)\|_{L^\infty} \le C\ \exp(-\min(a, \theta - kf(1))t)$, $t \ge 0$.

Now we consider the case $kf(1) = \theta$, $\beta > 0$. Integrating the v-equation in (1.5) gives

$$\bar{v}'(t) \le -\beta \int_0^1 v^2(x,t)dx \le -\beta\bar{v}^2(t),\ t \ge 0.$$

So

$$\bar{v}(t) \leq \frac{C}{\beta t + 1},\ t \geq 0. \tag{5.3}$$

Let $z(x,t) = (\beta t + 1)v(x,t)$. Then

$$\begin{cases} z_t \leq (\lambda z_x - \chi\phi'(u)u_x z)_x + \beta z, \\ \lambda z_x - \chi\phi'(u)u_x z = 0 \text{ at } \ x = 0, 1, \\ z(x,0) = v_0(x). \end{cases} \tag{5.4}$$

Again, (5.3) and the proof of Lemma 4.7 imply that $\|z(\cdot,t)\|_{L^\infty} \leq \text{const } C$, i.e., $\|v(\cdot,t)\|_{L^\infty} \leq C/(\beta t + 1),\ t \geq 0$. By a comparison argument as above, we also have $\|1 - u(\cdot,t)\|_{L^\infty} \leq C/(\beta t + 1)$, $t \geq 0$.

Next, we consider the case $kf(1) = \theta$, $\beta = 0$. Integrating the v-equation in (1.5) and using the fact that $kf(u) - \theta = kf(u) - kf(1) \leq -k \min_{\eta\in[0,1]} f'(\eta)\ (1-u)$, we have

$$\bar{v}'(t) \leq -C \int_0^1 (1 - u(x,t))v(x,t)dx,\ t \geq 0. \tag{5.5}$$

Because of the global boundedness of v, we have

$$\int^\infty \int_0^1 (1 - u(x,t))v(x,t)dx\, dt < \infty. \tag{5.6}$$

Let $w = 1 - u$. By the u-equation in (1.5), we obtain

$$\frac{1}{2}\frac{d}{dt}\int_0^1 w^2(x,t)dx \leq -hw^2(1,t) - \int_0^1 w_x^2(x,t)dx + C\int_0^1 w(x,t)v(x,t)dx.$$

This and (5.6) imply

$$\int^\infty w^2(1,t)dt < \infty,\ \int^\infty \int_0^1 w_x^2(x,t)dx\, dt < \infty. \tag{5.7}$$

By the second part of (iii) of Theorem 4.1, $w(1,t)$ and $\int_0^1 w_x^2(x,t)dx$ are uniformly continuous on $[0,\infty)$ and hence by (5.7), $w(1,t) \to 0$ and $\int_0^1 w_x^2(x,t)dx \to 0$ as $t \to \infty$. Now

$$|w(x,t) - w(1,t)| \leq \int_0^1 |w_x(x,t)|dx \leq \left(\int_0^1 w_x^2(x,t)dx\right)^{1/2} \to 0$$

as $t \to 0$. Thus $\|1 - u(\cdot,t)\|_{L^\infty} \to 0$ as $t \to \infty$.

We proceed to show $\|v(\cdot,t)\|_{L^\infty} \to 0$ as $t \to \infty$. Since $\bar{v}'(t) + k\bar{u}'(t) = kh(1 - u(1,t)) - \theta\bar{v}(t)$ and since $\bar{v}(t)$ is decreasing (see (5.5)), $\bar{v}(t) \to 0$ as $t \to \infty$. By the v-equation in (1.5) again,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 v^2(x,t)dx &\leq -\lambda\int_0^1 v_x^2(x,t)dx + \chi\int_0^1 \phi'(u)u_x v v_x dx \\ &\leq -\frac{\lambda}{2}\int_0^1 v_x^2(x,t)dx + C\int_0^1 u_x^2(x,t)dx.\end{aligned}$$

Then the global boundedness of v and (5.7) imply

$$\int^{\infty}\int_0^1 v_x^2(x,t)dx\,dt<\infty. \tag{5.8}$$

By the second part of (iii) of Theorem 4.1 again, $\int_0^1 v_x^2(x,t)dx$ is uniformly continuous on $[0,\infty)$ and hence by (5.8), we have $\int_0^1 v_x^2(x,t)dx\to 0$ as $t\to\infty$. Now

$$|v(x,t)-\bar{v}(t)|\leq\left(\int_0^1 v_x^2(x,t)dx\right)^{1/2}\to 0,$$

and thus $\|v(\cdot,t)\|_{L^\infty}\to 0$ as $t\to\infty$.

Proof of (ii) *of Theorem* 5.1. Let N_δ be the neighborhood of the trivial steady state $(1,0)$ consisting of (u,v) such that $\|1-u\|_{L^\infty}+\|v\|_{L^\infty}<\delta$. We want to show that for a small $\delta>0$, the solution $(u(\cdot,t),v(\cdot,t))$ of (1.5) always leaves N_δ in finite time no matter how close the initial value (u_0,v_0) is to $(1,0)$ (assuming (u_0,v_0) satisfies the condition in Theorem 4.8). If this is not true, then for a small $\delta>0$,

$$kf(u(x,t))-\theta-\beta v(x,t)\geq(kf(1)-\theta)/2,\ x\in[0,1],\ t\geq 0.$$

Thus by integrating the v-equation in (1.5), we have

$$\frac{d}{dt}\int_0^1 v(x,t)dx\geq\frac{(kf(1)-\theta)}{2}\int_0^1 v(x,t)dx,\ t\geq 0.$$

This obviously makes it impossible for $(u(\cdot,t),v(\cdot,t))$ to stay in N_δ forever. This completes the proof of Theorem 5.1.

Next we wish to establish the asymptotic stability of the positive steady states of (1.5) (again $\beta=0$ is allowed) when θ less than but close to $kf(1)$. By the proof of Theorem 3.1, (1.5) has a unique positive steady state $(u(s,x),v(s,x))$, $0<s<\varepsilon$, such that $kf(1)-\theta=\mu(s)$, $u(s,\cdot)=1-s(K_1f(1)+\psi_1(s))$, $v(s,\cdot)=s(1+\psi_2(s))$. Recall that $\mu(0)=0<\mu'(0)$, $\psi_1(0)=\psi_2(0)=0$.

THEOREM 5.2. *For θ less than but close to $kf(1)$, the unique positive solution of* (1.5) *($\beta=0$ allowed) is exponentially asymptotically stable in the $H^1(0,1)$ topology.*

Proof. Linearize (1.5) at $(u(s,\cdot),v(s,\cdot))$. By the principle of the linearized stability ([S, Thm. 8.6], [D, Thm. 5.2]), we only need to show the negativeness of the real part of the eigenvalues η of the linearized elliptic problem

$$\begin{cases} u''-f'(u(s,\cdot))v(s,\cdot)u-f(u(s,\cdot))v=\eta u,\\ [\lambda v'-\chi\phi''(u(s,\cdot))u'(s,\cdot)v(s,\cdot)u-\chi\phi'(u(s,\cdot))v(s,\cdot)u'\\ \qquad -\chi\phi'(u(s,\cdot))u'(s,\cdot)v]'+(kf(u(s,\cdot))-kf(1)-\beta v(s,\cdot))v\\ \qquad +(kf'(u(s,\cdot))u-\beta v)v(s,\cdot)+\mu(s)v=\eta v,\\ u'(0)=0=u'(1)+hu(1),\\ \lambda v'-\chi\phi''(u(s,x))u'(s,x)v(s,x)u-\chi\phi'(u(s,x))v(s,x)u'\\ \qquad -\chi\phi'(u(s,x))u'(s,x)v=0\quad\text{at}\quad x=0,1. \end{cases} \tag{5.9}$$

Multiplying u^* and v^* (the conjugates of u and v) to the u-equation and v-equation in (5.9), respectively, then integrating by parts, we see that (a) the real part of the eigenvalues is bounded above uniformly with respect to small $s>0$; (b) the imaginary part of the eigenvalues in any fixed vertical strip on the plane is bounded uniformly with respect to small $s > 0$.

Then since the set of the eigenvalues is discrete, there is an eigenvalue, denoted by $\eta(s)$, whose real part is the largest. We wish to show $\operatorname{Re}\eta(s) < 0$ for all small $s > 0$. Suppose otherwise; then $\operatorname{Re}\eta(s) \geq 0$ along a subsequence $s \to 0$. Now by the above discussion, $\eta(s)$ is bounded and hence we can assume $\eta(s) \to \eta_0$ as $s \to 0$, where $\operatorname{Re}\eta_0 \geq 0$.

We claim $\eta_0 = 0$. Denote the eigenvector corresponding to $\eta(s)$ by $(\widetilde{u}(s,\cdot), \widetilde{v}(s,\cdot))$ which is normalized: $\|\widetilde{u}(s,\cdot)\|^2_{L^2(0,1)} + \|\widetilde{v}(s,\cdot)\|^2_{L^2(0,1)} = 1$. By the elliptic regularity theory, after passing to a subsequence, $(\widetilde{u}(s), \widetilde{v}(s)) \to (\widetilde{u}_0, \widetilde{v}_0)$ in $C^2([0,1])$ as $s \to 0$. $(\widetilde{u}_0, \widetilde{v}_0)$ is nonzero and satisfies

$$\begin{cases} \widetilde{u}_0'' - f(1)\widetilde{v}_0 = \eta_0 \widetilde{u}_0, \\ \lambda \widetilde{v}_0'' = \eta_0 \widetilde{v}_0, \\ \widetilde{u}_0'(0) = 0 = \widetilde{u}_0'(1) + h\widetilde{u}_0(1), \\ \widetilde{v}_0'(0) = 0 = \widetilde{v}_0'(1). \end{cases}$$

Thus $\eta_0 = 0$ and $(\widetilde{u}_0, \widetilde{v}_0) = (-K_1 f(1), 1)/(1 + \|K_1 f(1)\|^2_{L^2})$.

We now renormalize $(\widetilde{u}(s), \widetilde{v}(s))$ so that it converges to $(-K_1 f(1), 1)$ in $C^2([0,1])$ as $s \to 0$. Since $\int_0^1 (kf(u(s,x)) - kf(1) - \beta v(s,x) + \mu(s))v(s,x)dx = 0$, we have ("$\bullet$" meaning differentiation in s-variable),

$$\begin{aligned} &\int_0^1 (kf'(u(s,x))\overset{\bullet}{u}(s,x) - \beta \overset{\bullet}{v}(s,x) + \mu'(s))v(s,x)dx \\ &\qquad + \int_0^1 (kf(u(s,x)) - kf(1) - \beta v(s,x) + \mu(s))\overset{\bullet}{v}(s,x)dx = 0. \end{aligned} \tag{5.10}$$

Integrating the v-equation in (5.9) (with $\eta = \eta(s)$, $(u,v) = (\widetilde{u}(s), \widetilde{v}(s))$, we have

$$\begin{aligned} &\int_0^1 (kf'(u(s,x))\widetilde{u}(s,x) - \beta \widetilde{v}(s,x))v(s,x)dx \\ &\qquad + \int_0^1 (kf(u(s,x)) - kf(1) + \mu(s) - \beta v(s,x))\widetilde{v}(s,x)dx \\ &= \eta(s) \int_0^1 \widetilde{v}(s,x)dx. \end{aligned} \tag{5.11}$$

Comparing (5.11) with (5.10), we have

$$\eta(s)(1 + o(1)) = o(s) - \mu'(s)s(1 + o(1)) = s(-\mu'(0) + o(1)).$$

Since $\mu'(0) > 0$, $\operatorname{Re}\eta(s) < 0$ for small $s > 0$, contradicting our assumption $\operatorname{Re}\eta(s) \geq 0$. Theorem 5.2 is proved.

REFERENCES

[A] N. Alikakos, *L^p bounds of solutions of reaction-diffusion equations*, Comm. Partial Differential Equations, 4 (1979), pp. 827–868.

[Am 1] H. Amann, *Dynamic theory of quasilinear parabolic equations* II: *Reaction-diffusion systems*, Differential Integral Equations, 3 (1990), pp. 13–75.

[Am 2] H. Amann, *Dynamic theory of quasilinear parabolic systems* III: *Global existence*, Math. Z., 202 (1989), pp. 219–250.

[BB] J. Blat and K. J. Brown, *Global bifurcation of positive solutions in some systems of elliptic equations*, SIAM J. Math. Anal., 17 (1986), pp. 1339–1353.

[CR] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[D] A.-K. Drangeid, *The principle of linearized stability for quasilinear parabolic evolution equations*, Nonlinear Anal., 13 (1989), pp. 1091–1113.

[H] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, Heidelberg, New York, 1981.

[KS] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.

[LSU] O. A. Ladyzenskaya, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, American Mathematical Society, Providence, RI, 1967.

[LAK] D. Lauffenburger, R. Aris, and K. Keller, *Effects of cell motility and chemotaxis on microbial population growth*, Biophys. J., 40 (1982), pp. 209–219.

[PW] W. K. Pilgram and F. D. Williams, *Survival value of chemotaxis in mixed cultures*, Canad. J. Microbiol., 22 (1976), pp. 1771–1773.

[R] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[S] G. Simonett, *Center manifolds for quasilinear reaction-diffusion systems*, Differential Integral Equations, 8 (1995), pp. 753–796.

[Z] B. Zeng, *Steady states of a chemotaxis system*, Appl. Math., 1 (1990), pp. 78–83 (in Chinese).

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 561–573

STOPPING TIMES AND LOCAL CONVERGENCE FOR SPLINE WAVELET EXPANSIONS[*]

RICHARD F. GUNDY[†] AND KAZAROS KAZARIAN[‡]

Abstract. A local convergence theorem for spline wavelet expansions is proved. This theorem relates the finiteness of the quadratic variation of the expansion with the local convergence of the expansion on sets of positive measure. A stability property of these expansions is one of the key points in the proof.

Key words. wavelets, stopping times, martingales

AMS subject classifications. 42C15, 60G46

PII. S0036141097327392

1. Introduction.

The principal purpose of this paper is to prove a local convergence theorem for spline wavelet expansions, using a combination of techniques from martingale theory and wavelet analysis. In particular, we show that the notion of a stopping time may be adapted to these wavelet expansions.

The Haar functions are the point of departure for this discussion. The Haar series may be viewed as the sequence of partial sums of orthogonal elements from the multiresolution analysis generated by the dilations and translations of $\chi_{[0,1]}(x)$, the indicator function of the unit interval. On the other hand, this sequence of partial sums forms a martingale with respect to the filtration of σ-fields generated by the (dyadic) dilations and (integer) translations of $\chi_{[0,1]}$. This coincidence is unique: no other multiresolution analysis generates martingales in a similar fashion. The following theorem concerning Haar functions is proved by martingale methods [3] (see also [5]).

THEOREM A. *Let $f = (f_0, f_1, \ldots)$ be the sequence of partial sums of a Haar series on $[0,1]$. Then the following sets are equivalent, i.e., they differ by at most a set of Lebesgue measure zero:*

$$A = \{x : f^*(x) := \sup_n |f_n(x)| < \infty\};$$

$$B = \left\{ x : S^2(f)(x) := \sum_0^\infty (f_n(x) - f_{n-1}(x))^2 < \infty \right\};$$

$$C = \{x : \limsup f_n(x) = \liminf f_n(x)\}.$$

This theorem is not true for other martingales, in general. Some additional stability condition must be assumed. The goal of this paper is to extend this local convergence theorem for a class of wavelet expansions satisfying a stability condition. This class includes the series arising from polynomial spline wavelets. The precise statement of the theorem, together with the stability condition, is given below.

[*] Received by the editors September 18, 1997; accepted for publication (in revised form) May 19, 1999; published electronically February 24, 2000.

http://www.siam.org/journals/sima/31-3/32739.html

[†] Department of Statistics, Rutgers University, Hill Center, Busch Campus, Piscataway, NJ 08855 (gundy@rci.rutgers.edu). This author's research was supported by the sabbatical program at the Universidad Autonoma de Madrid during summer 1994.

[‡] Department of Mathematics, Universidad Autonoma de Madrid, Madrid, Spain (kazaros.kazarian@uam.es). This author's research was supported by DGICYT grants PB94-0149 and PB97-0030.

2. Notation. We will use a *multiresolution analysis* $V = \{V_j; -\infty < j < \infty\}$ generated by a compactly supported, continuous (pre)scale function $\phi(x)$. That is, we require that ϕ have a nonzero integral, that it satisfy a dilation equation of the form

$$\phi(x) = \sum_{k=0}^{N} p_k \phi(2x - k), \tag{a}$$

and that

$$\sum a_n^2 \cong \left\| \sum a_k \phi(x-k) \right\|_2^2 \tag{b}$$

(the translates of ϕ are a Riesz basis). Linear combinations of the translates of ϕ form a subspace of $L^2(\mathbf{R})$; the L^2-closure of this subspace is denoted by V_0. Dyadic dilations of functions in V_0 form a closed subspace of $L^2(\mathbf{R})$, denoted by V_j ($f(\cdot) \in V_j$ iff $f(2^{-j}\,\cdot) \in V_0$). If $V_j \subset V_{j+1}$ for all integers j it then follows that the increasing sequence of subspaces $\{V_j\}$ exhausts $L^2(\mathbf{R})$ (see [6, Chapter 2, p. 48]). Let P_j be the orthogonal projection operator from L^2 onto V_j.

A special case of a multiresolution analysis of this type is the one generated by $\phi(x) = \chi_{[0,1]}(x)$. In this case, the projections P_j are conditional expectations, and a sequence $P_j(f)(x)$ forms a martingale. To distinguish this special case, we use D_0 to denote the set of *all* linear combinations of translates $\chi_{[0,1]}(\cdot - k)$, $k \in \mathbf{Z}$, and $D = \{D_j; -\infty < j < \infty\}$ to denote the set of 2^j-dilates of elements of $D_0 : D_j = \{f : f(2^{-j}x) \in D_0\}$. A *stopping time* $\tau(x)$ is a function with values in $\mathbf{Z}_+ \cup \{\infty\}$, such that the indicator function of $\{x : \tau(x) = j\}$ belongs to D_j.

3. Prewavelets with compact support. In this paper we assume that ϕ is supported in the interval $[0, N]$. This implies the existence of a prewavelet ψ having compact support. By this, we mean a function $\psi \in V_0$ such that $\psi(2x - 2k)$, $k \in \mathbf{Z}$, is a Riesz basis in the orthogonal complement of V_0 in V_1, usually written as W_0. The function $\psi(2\,\cdot)$ is necessarily of the form

$$\psi(2x) = \sum_{-N+1}^{2N-1} c_k \phi(2x - k),$$

and since $\psi(2\,\cdot) \in W_0$, we have

$$\int \phi(x)\bar{\psi}(2x)dx = \sum_{-N+1}^{2N-1} \bar{c}_k \int \phi(x)\overline{\phi(2x-k)}dx = 0.$$

If we set

$$e_k = \int \phi(x)\overline{\phi(2x-k)}dx,$$

one solution to this equation is given by $c_k = (-1)^k \bar{e}_{1-k}$; the products $\bar{c}_k e_k$ and $\bar{c}_{1-k} e_{1-k}$ have opposite signs and equal magnitudes and so cancel in pairs. The function $\psi(2x)$ is compactly supported on the interval $[-N+1, 3N-1]$ since $e_k \equiv 0$ if $k \notin [-N+1, 2N-1]$. It turns out that, with this choice, the translates $\psi(2x - 2k)$, $k \in \mathbf{Z}$, form a Riesz basis for W_0 (see [6]). (We pause here to note that our notation is slightly different from the usual custom, where $\psi(x)$ is taken to be a function in W_0, whose translates $\psi(x-k)$, $k \in \mathbf{Z}$, form a Riesz basis. Our notation seems to us to make the exposition run more smoothly.) We shall also use the notions of dual prescale function and dual prewavelet (see Chui [2, Chapter 5, p. 170]). The dual prescale function $\tilde{\phi}$ is a function that belongs to V_0 and satisfies

$$\int \phi(x-k)\overline{\tilde{\phi}(x)}dx = \delta_0(k).$$

The dual prewavelet satisfies $\tilde{\psi}(2x) \in W_0$:

$$\int \psi(2x-2k)\overline{\tilde{\psi}(2x)}dx = \delta_0(k).$$

We now impose an additional stability condition on the prescale function ϕ.

Condition (M-Z). Given any measurable subset $E \subseteq [0,1]$, of measure greater than $\delta > 0$, and any sequence a_k, $k = -N+1, \ldots, 0$, we have

$$\sum_{k=-N+1}^{0} |a_k| \le c(\delta) \sup_{x\in E} \left| \sum_{k=-N+1}^{0} a_k \phi(x-k) \right|,$$

where $c(\delta)$ depends on δ but is otherwise independent of E.

Remarks. Since ϕ is assumed to be bounded, the inequality may be reversed at the expense of another constant. That is,

$$c^{-1}(\delta) \sum_{k=-N+1}^{0} |a_k| \le \sup_{x\in E} \left| \sum_{k=-N+1}^{0} a_k \phi(x-k) \right| \le c \sum_{k=-N+1}^{0} |a_k|.$$

This "two-norm" condition is in the same spirit as a stability condition suggested by Marcinkiewicz and Zygmund in their study of convergence of series of independent random variables [8]. Their stability condition was introduced in a conditional form in [4] to prove a generalization of Theorem A. This conditional stability condition is the basis for the results in [1].

The Haar scale function $\chi_{[0,1]}(x)$ satisfies Condition (M-Z) in a vacuous way. On the other hand, if we take the N-fold convolution of $\chi_{[0,1]}$ with itself, we obtain a prescale function ϕ for the multiresolution analysis of polynomial splines of degree $N-1$. The translates $\phi(x-k)$, $k = -N+1, \ldots, 0$, are linearly independent polynomials on $[0,1]$. Given $E \subseteq [0,1]$ of measure greater than $\delta > 0$, we can find $x_1, \ldots, x_N$, $x_i \in E$, such that

$$\min_{i,j} |x_i - x_j| = O_N(\delta).$$

Now write

$$\sum_{-N+1}^{0} a_k \phi(x-k) = \sum_{k=0}^{N-1} b_k x^k$$

and compute the coefficients b_k by using Cramer's rule on the Vandermonde determinant at the points $x_1, \ldots, x_N$. This gives an estimate for the b-sequence:

$$\sum_{k=0}^{N-1} |b_k| \le c(\delta) \sup_{x\in E} \left| \sum_{k=-N+1}^{0} a_k \phi(x-k) \right|,$$

and consequently, the same estimate for the a-sequence:

$$\sum_{k=-N+1}^{0} |a_k| \leq c'(\delta) \sup \left| \sum_{k=-N+1}^{0} a_k \phi(x-k) \right| .$$

Results. The principal result is as follows.

THEOREM B. *Suppose that ϕ is a prescale function that satisfies Condition* (M-Z). *Suppose that $f = (f_{j-1}, f_j, \ldots)$ is a sequence of functions such that*

$$P_j(f_{j+1}) = f_j$$

for every $j \in \mathbf{Z}$. Then the following sets are equal almost everywhere (a.e.):

$$A = \left\{ x : f^*(x) := \sup_j |f_j(x)| < \infty \right\};$$

$$B = \left\{ x : S^2(f)(x) = \sum_{j=-\infty}^{\infty} (f_j(x) - f_{j-1}(x))^2 < \infty \right\};$$

$$C = \left\{ x : \overline{\lim} f_j(x) = \underline{\lim} f_j(x) < \infty \right\}.$$

(That is, the Lebesgue measure of the sets $A \triangle B, A \triangle C, C \triangle B$ vanishes.)

Proof. Before giving the details of the proof, let us outline the strategy. It is enough to prove the theorem with the sets A, B, C restricted to any interval of unit length. Therefore, we assume that A, B, and C are subsets of the unit interval. Given $\epsilon > 0$, we may also assume that the measure of any one of these sets is greater than $1 - \epsilon$. Both of these assumptions amount to an appropriate choice of an origin for the dilation scale. Since $f_0 \in L^2(\mathbf{R})$, we have

$$\sup_{j \leq 0} |f_j(x)| < \infty;$$

$$\sum_{j \leq 0} (f_j(x) - f_{j-1}(x))^2 < \infty.$$

(See [6].) Therefore, we restrict attention to $f_j(x)$, $x \in [0,1]$, and $j \geq 0$. □

The basic idea of the proof is to introduce a (dyadic) stopping time $\tau = \tau_\lambda$ with the property that $\{\tau(x) = \infty\}$ (essentially) coincides with the set A_λ (or B_λ) where $f^*(x) \leq \lambda$ (or $S(f) < \lambda$). Furthermore, the stopping time should be such that the stopped sequence $f_{\tau(\cdot) \wedge j}(\cdot)$ is uniformly bounded in $L^2(\mathbf{R})$. *If it were true* that the projections P_j commute with the stopping time, i.e.,

$$P_j(f_{\tau \wedge (j+1)}) = f_{\tau \wedge j},$$

then the sequence $f^\tau = (f_{\tau \wedge 1}, f_{\tau \wedge 2}, \ldots)$ would satisfy

$$S^2(f)(x) = S^2(f^\tau)(x)$$

on $\{x : \tau(x) = \infty\}$. This would imply that $A_\lambda \subseteq B$ since

$$\|S(f^\tau)\|_2 = \sup_j \|f_{\tau \wedge j}\|_2 < \infty.$$

The commutation hypothesis (that P_j commute with all dyadic stopping times) is equivalent to assuming that the P_j are dyadic conditional expectations. In this case, the sequence f_j is forced to be a martingale. Since those sequences we deal with here are not martingales, we cannot prove the theorem in this way. Nevertheless, the strategy may be rescued from obvious failure. The proof is carried out by an inductive procedure that involves a number of details. In order to clarify the description, we proceed in steps.

Step 1. We now prove that $A \subseteq B$, assuming that $A \subseteq [0,1]$ and that $\text{meas}(A) > 1-\epsilon$. Here $\epsilon > 0$ is assumed to be small, subject to constraints that will become clear as the discussion unfolds. Since $f_0(x)$ is restricted to the unit interval, we may write

$$f_0(x) = \sum_{j=-N+1}^{0} a_j \phi(x-j).$$

Choose $\lambda > 0$ and define

$$A_\lambda^c := \left\{ x : \sup_{j \geq 0} |f_j(x)| > \lambda \right\}.$$

Then A_λ^c satisfies $A_\lambda^c \supset A^c$, and if λ is large enough, we will have

$$\text{meas}(A_\lambda^c) \leq 2\epsilon.$$

We construct a covering of A_λ^c by dyadic intervals, defined by means of a stopping time $\gamma(x)$ relative to the dyadic multiresolution analysis D. This stopping time will be modified by subsequent considerations before we are finished.

Step 2. We wish to project the characteristic function of the set A_λ^c onto the space D_j; let g_j be this projection. The function g_j is just the average of the characteristic function of A_λ^c over each dyadic interval of length 2^{-j}. The sequence g_j, $j \geq 0$, is a dyadic martingale. Define

$$\gamma(x) = \inf \{ j : g_j(x) \geq 1/2 \}, \quad \text{where } \gamma(x) = \infty \text{ if this set is empty.}$$

By Doob's maximal inequality, applied to the sequence g_j, $j \geq 0$, we have

$$\text{meas}\{x : \gamma(x) < \infty\} \leq 2 \ \text{meas}(A_\lambda^c) \leq 4\epsilon.$$

The stopping time $\gamma(x)$ determines a family of disjoint subintervals of $[0,1]$ defined by $\{x : \gamma(x) = j\}$. These intervals are unions of dyadic intervals of length 2^{-j}. A connected union of maximal length is called a component of $\{x : \gamma(x) = j\}$.

Step 3. Starting with level $n_1 = \inf_x \gamma(x)$, the components of $\{x : \gamma(x) = n_1\}$ are divided into three classes: The first class consists of all components of length less than $N \cdot 2^{-n_1}$. The second class consists of all components of length greater than or equal to $N \cdot 2^{-n_1}$ but less than $8N \cdot 2^{-n_1}$. The third class consists of components of length greater than or equal to $8 \cdot N \cdot 2^{-n_1}$. We modify the intervals of the second class by extending them, to the right, by a union of dyadic intervals such that the total length of the enlarged interval (the component plus the added intervals) equals $8N \cdot 2^{-n_1}$. The entire set of intervals obtained in this way may again be divided into two components: (a) short components of length less than $N \cdot 2^{-n_1}$ and (b) the remaining components of length greater than or equal to $8N \cdot 2^{-n_1}$. We modify the stopping time $\gamma(x)$ to obtain another one, $\tau(x)$, in the following ways: (a) $\tau(x) = n_1$

for points in any component containing the point $x = 0$ or $x = 1$ if such a component exists. (b) $\tau(x) = n_1$ for points in the long components (of length greater than or equal to $8 \cdot N2^{-n_1}$). (c) On the remaining points, $\tau(x) > n_1$ and will be defined by the inductive process. To recapitulate, we have enlarged the set $\{x : \gamma(x) = n_1\}$ to obtain a new set of components. The measure of the enlarged set is no greater than nine times the measure of the components of $\{x : \gamma(x) = n_1\}$ of intermediate length. The stopping time $\tau(x) = n_1$ on the "long" components and on the components, if they exist, containing $x = 0$ or $x = 1$. The "short" components of the enlarged set are those that do not contain $x = 0$ or $x = 1$ and whose length does not exceed $N \cdot 2^{-n_1}$.

Step 4. We now continue the induction as follows: Consider the set of short components, just cited, together with the set $\{x : \gamma(x) = n_1 + 1\}$. We combine the short components of the set $\{x : \gamma(x) = n_1\}$ and those components of the set $\{x : \gamma(x) = n_1 + 1\}$ that do not belong to the set $\{x : \tau(x) = n_1\}$. The union of these two sets is a collection of dyadic intervals of length $2^{-(n_1+1)}$. The components of this set are sorted into three categories as before: those of length less than $N \cdot 2^{-(n_1+1)}$, those of intermediate length, from $N \cdot 2^{-(n_1+1)}$ to $8N \cdot 2^{-(n_1+1)}$, and those of length greater than or equal to $8N \cdot 2^{-(n_1+1)}$. The components of intermediate length are enlarged so that they have length exactly $8N \cdot 2^{-(n_1+1)}$. (This enlargement is made simply by adjoining adjacent dyadic intervals, of length $2^{-(n_1+1)}$, to the right of the component in question. The new component may overlap some of the set $\{x : \tau(x) = n_1\}$, as well as other components of $\{x : \gamma(x) \leq n_1 + 1\}$ not contained in $\{x : \tau(x) = n_1\}$.) We now combine the short, enlarged, and long components of this set *together with* the components of the set $\{x : \tau(x) = n_1\}$. Since the components of $\{x : \tau(x) = n_1\}$ are all longer than $8N \cdot 2^{-n_1}$, the *components of the combined set* are either longer than or equal to $8 \cdot N \cdot 2^{-(n_1+1)}$ or shorter than $N \cdot 2^{-(n_1+1)}$.

We now define $\tau(x) = n_1 + 1$ for points x in a long component of the combined set if $\tau(x)$ has not been previously defined. We also define $\tau(x) = n_1 + 1$ if x is in a component (long or short) that contains 0 or 1 and $\tau(x)$ has not been previously defined. Thus, the only components that remain are short and isolated; that is, they are of length less than $N \cdot 2^{-(n_1+1)}$ and lie in the interior of the *complement* of the long components and contain neither 0 nor 1.

Step 5. The passage from n_1 to n_1+1 is indicative of the induction procedure. At the nth stage, the set $\{x : \tau(x) = n\}$ has been defined. The short components defined by the procedure are of length less than $N \cdot 2^{-n}$ and lie in the interior of the unit interval, separated from the long components by a distance of at least 2^{-n}. These short components are a part of the set $\{x : \gamma(x) \leq n\}$.

The short components are combined with that part of the set $\{x : \gamma(x) = n + 1\}$ that remains in the complement of $\{x : \tau(x) \leq n\}$ and are treated as in Step 3 to define the set $\{x : \tau(x) = n + 1\}$.

Step 6. We now make a penultimate adjustment in the definition of $\tau(x)$. At each stage, if the set $\{x : \tau(x) = n\}$ produces complementary intervals belonging to the complement of $\{x : \tau(x) \leq n\}$ that are of length less than $8N \cdot 2^{-n}$, we adjoin them to the set $\{x : \tau(x) = n\}$. (That is, we define $\tau(x) = n$ on these intervals also.) This addition will take place only if new points were added to the set $\{x : \tau(x) < n\}$ by additional stopping. A fixed component of $\{x : \tau(x) = n\}$ can have at most two "small" complementary contiguous intervals, and the components of $\{x : \tau(x) = n\}$ are at least of length 2^{-n}. Therefore, this addition multiplies the measure of $\{x : \tau(x) = n\}$ by a factor no greater than $16N$.

Step 7. It remains for us to estimate the measure of the set $\{x : \tau(x) < \infty\}$. First of all, let us observe that $\tau(x)$ satisfies

$$\gamma(x) \leq \tau(x) \leq \gamma(x) + \log_2 N$$

since any component of $\{x : \gamma(x) = n\}$ must be "long" or "intermediate" if the scale is 2^{-m}, where $N \cdot 2^{-m} \leq 2^{-n}$. Finally, the enlargement procedure has been done in such a way that

$$\operatorname{meas}\{x : \tau(x) < \infty\} \leq 9 \operatorname{meas}\{x : \gamma(x) < \infty\} = O(\epsilon).$$

Step 8. Now let us give a preliminary definition of the stopped sequence $\tilde{f}^{\tau}_j$. As we pointed out above, the straightforward approach, where $\tilde{f}^{\tau}_j(x) := f_{j\wedge\tau(x)}(x)$, is not suitable. However, this procedure may be modified as follows: The sequence $f_0, f_1, \ldots, f_{n_1-1}$ is not altered. For the index n_1, we consider the intervals that are *complementary* to $\{x : \tau(x) = n_1\}$, that is, the set $\{x : \tau(x) > n_1\}$. A typical complementary interval (in $\{x : \tau(x) > n_1\}$) consists of a union of dyadic intervals of length 2^{-n_1}. Each of these dyadic intervals may be classified by the stopping time γ: on a given interval of length 2^{-n_1}, $\gamma(\cdot) \equiv n_1$ or $\gamma(\cdot) > n_1$. By construction, the extreme left and right dyadic intervals (of the entire complementary interval) are of the latter type, where $\gamma(\cdot) > n_1$. The dyadic intervals in the interior may be of either type. However, the short components of $\{x : \gamma(x) = n_1\}$ are of length less than $N \cdot 2^{-n_1}$. The complementary interval may, of course, contain many such components, separated from each other by dyadic intervals where $\gamma(\cdot) > n_1$. Finally, the complementary interval contains neither endpoint $x = 0$ nor $x = 1$.

On a fixed complementary interval the function $f_{n_1}(x)$ may be represented as a finite sum $\tilde{f}^{\tau}_{n_1}$,

$$\tilde{f}^{\tau}_{n_1}(x) := \sum_{k=\ell}^{r} a_k \phi(2^{n_1}x - k).$$

Here we assume that the complementary interval has left endpoint $(\ell + N - 1) \cdot 2^{-n_1}$ and right endpoint $(r + 1) \cdot 2^{-n_1}$, so that the above is the shortest representation of $f_{n_1}(x)$ on the interval. The sum does not necessarily give the value of $f_{n_1}(x)$ outside the complementary interval, nor are we assured that the difference

$$\begin{aligned} \tilde{d}^{\tau}_{n_1}(x) &= \tilde{f}^{\tau}_{n_1}(x) - \tilde{f}_{n_1-1}(x) \\ &\quad (= \tilde{f}^{\tau}_{n_1}(x) - f_{n_1-1}(x)) \end{aligned}$$

belongs to the space W_{n_1-1}. However, we can assert that

$$\max_{k \leq n_1} |\tilde{f}^{\tau}_k(x)| \leq O(\lambda)$$

for all $x \in \mathbf{R}$. This is true because of Condition (M-Z). The argument is as follows: For $\tilde{f}^{\tau}_k = f_k$, $k < n_1$, on *every* dyadic interval of length 2^{-k}, the (closed) subset of points $\{x : \sup_{j\geq 0} |f_j(x)| \leq \lambda\}$ contained in the dyadic interval has a proportion that is less than $1/2$. Condition (M-Z) then guarantees that

$$|f_k(x)| = O(\lambda)$$

uniformly in the dyadic interval. (See the Remarks after the statement of Condition (M-Z).) The argument for $\tilde{f}^{\tau}_{n_1}$ ($= f_{n_1}$ on the complementary interval) is similar but a bit more complicated. The function $|\tilde{f}^{\tau}_{n_1}(x)| = O(\lambda)$ on any interval where $\gamma(x) > n_1$ for the reason just stated. On the stretches (short components) of length less than $N \cdot 2^{-n_1}$, where $\gamma(x) = n_1$, we cannot apply this argument directly. However, the values of $\tilde{f}^{\tau}_{n_1}(\cdot)$ in this stretch are majorized by a constant times the sum of the moduli of all coefficients a_k that enter into the representation

$$\tilde{f}^{\tau}_{n_1}(x) = \sum_{k=\ell}^{r} a_k \phi(2^{n_1}x - k)$$

on this stretch. Because the stretch is short, the translates $\{k\}$ specific to this stretch also appear in the representation of the function $\tilde{f}^{\tau}$ on the dyadic intervals of length 2^{-n_1}, I_0, and I_1, that bound the stretch. (Some translates are associated with I_0, some are associated with I_1.) On each interval I_i, $i = 0, 1$, we know that $\gamma(\cdot) > n_1$, so that $|\tilde{f}^{\tau}_{n_1}(x)| = O(\lambda)$ on these intervals. By Condition (M-Z), the sum of the moduli of the corresponding coefficients is of the same order. On the stretch, $|\tilde{f}^{\tau}_{n_1}(x)|$ is also majorized, up to a constant, by the sum of all of these coefficients. Therefore, we may conclude that $|\tilde{f}^{\tau}_{n_1}(x)| = O(\lambda)$ on the entire complementary interval. The constants depend only on N, the bound on $|\phi(x)|$, and the constants from Condition (M-Z).

Outside the complementary interval, we claim that $\tilde{f}^{\tau}_{n_1}(x)$ satisfies the same estimate. In fact, the support of any sum

$$\tilde{f}^{\tau}_{n_1}(x) = \sum_{k=\ell}^{r} a_k \phi(2^{n_1}x - k)$$

contained in an interval consisting of the complementary interval, together with two intervals, to the left and right, of the complementary interval. Since $\phi(2^{n_1}x)$ has support on $[0, N \cdot 2^{-n_1}]$, the additional intervals are at most of length $(N-1)2^{-n_1}$. However, we know that the intervals where $\tau(x) = n_1$ are of length at least $8N \cdot 2^{-n_1}$, so the supports of the sums defining $\tilde{f}^{\tau}_{n_1}$ are disjoint. Since $|\tilde{f}_{n_1}(x)| = O(\lambda)$ on each complementary interval, Condition (M-Z) implies that this estimate holds on the entire support. Therefore, $|\tilde{d}^{\tau}_{n_1}| = O(\lambda)$ also.

Step 9. As noted above, the difference

$$\tilde{d}^{\tau}_{n_1}(x) = \tilde{f}^{\tau}_{n_1}(x) - \tilde{f}^{\tau}_{n_1-1}(x)$$

belongs to V_{n_1}, but not necessarily to W_{n_1-1}, the orthogonal complement of V_{n_1-1} in V_{n_1}, even though it represents d_{n_1} on each complementary interval. To remedy this, our strategy will be to obtain an expression for the difference $\tilde{d}^{\tau}_{n_1}(x)$ in terms of the prewavelets $\psi(2^{n_1}x - 2k)$ and to estimate the magnitude of the new representation. The stopping time τ will be altered again to obtain another stopping time ρ, such that the measure of $\{x : \rho(x) < \infty\}$ is larger than that of $\{x : \tau(x) < \infty\}$ by a fixed multiple. With this stopping time, the difference $f^{\rho}_{n_1} - f^{\rho}_{n_1-1}$ will be contained in W_{n_1-1}. The procedure will then be carried out for $n \geq n_1$, and we will be able to estimate the quantity $\sup_n \|f^{\rho}_n\|_2$.

On each complementary interval, $\tilde{d}^{\tau}_{n_1}(x)$ has a representation

$$\tilde{d}^{\tau}_{n_1}(x) = \sum_{k=\ell}^{r} e_k \phi(2^{n_1}x - k)$$

with possibly different ℓ, r. The supports of these sums are disjoint for the reasons cited above. Each sum composing $\tilde{d}^{\tau}_{n_1}$ may be restricted to a function that belongs to W_{n_1-1}. In fact, the original difference

$$\begin{aligned} d_{n_1}(x) &= f_{n_1}(x) - f_{n_1-1}(x) \\ &= \sum_s b_{2s}\psi(2^{n_1}x - 2s) \end{aligned}$$

belongs to W_{n_1-1} by assumption. We define $d_n^\tau(x)$ as the sum

$$d^\tau_{n_1}(x) = \sum_m b_{2m}\psi(2^{n_1}x - 2m),$$

where the indicated sum is taken over indices $2m$, $\ell + N - 1 \leq 2m \leq r - (2N-1)$. With this definition, $d^\tau_{n_1}$ belongs to W_{n_1-1} and its support is contained in the support of $\tilde{d}^\tau_{n_1}$.

(a) Since each complementary interval is at least of length $8N\cdot 2^{-n_1}$ $(r-\ell \geq 9N-1)$ and the support of $\psi(2^{n_1}x)$ is contained in an interval of length $(4N-2)2^{-n_1}$, the sums are nonempty.

(b) Furthermore, the supports of the various sums corresponding to disjoint complementary intervals are disjoint, because of the support remark above.

(c) The coefficients $|b_{2m}| = O(\lambda)$. In fact,

$$b_{2m} = \int d^\tau_{n_1}(x)\overline{\tilde{\psi}(2^{n_1}x - 2m)}dx,$$

where $\tilde{\psi}(2^{n_1}x - 2m)$ is the dual wavelet, acting as a linear functional of $W_{n_1-1} \subset V_{n_1}$. As a function on V_{n_1}, $\tilde{\psi}(2^{n_1}x - 2m)$ may be expressed as a linear combination of translates of the dual prescale function $\tilde{\phi}$. To estimate b_{2m}, it is enough to restrict the translates $\tilde{\phi}(2^{n_1}x - k)$ to those indices k such that $\phi(2^{n_1}x - k)$ is contained in the support of $\psi(2^{n_1}x - 2m)$. There are at most $3N-1$ such translates, and we have

$$\begin{aligned} |e_k| &= \left| \int d^\tau_{n_1}(x)\tilde{\phi}(2^{n_1}x - k)dx \right| \\ &= O(\lambda). \end{aligned}$$

Therefore, $|b_{2m}| = O(\lambda)$ also, since b_{2m} is a linear combination of at most $3N-1$ coefficients e_k.

(d) We recall that, on a fixed complementary interval,

$$\tilde{d}^\tau_{n_1}(x) = \sum_{k=\ell}^{r} e_k\phi(2^{n_1}x - k)$$

represents the difference $f^\tau_{n_1} - f^\tau_{n_1-1}$ on the complementary interval, but not necessarily outside this interval, where $\tilde{d}^\tau_{n_1}(x)$ has its support. The total support of $\sum_{k=\ell}^{r} e_k\phi(2^{n_1}x - k)$ consists of two intervals of length $(N-1)2^{-n_1}$, to the left and right of the complementary interval. Because the representation of

$$d^\tau_{n_1}(x) = \sum b_{2m}\psi(2^{n_1}x - 2m)$$

involves the function $\psi(2^{n_1}x)$, whose support is an interval of length $(4N-2)2^{-n_1}$, we need $(2N-1)$ translates to represent a function on intervals of length $2\cdot 2^{-n_1}$.

On the other hand, in order to obtain the estimate $|b_{2m}| = O(\lambda)$ we are only allowed translates $2m \leq r - (2N-1)$. All of this means that the sum

$$d^\tau_{n_1}(x) = \sum_{k=\ell}^{r} b_{2m}\psi(2^{n_1}x - 2m)$$

represents $d_{n_1}(x)$ for all x in the complementary interval except possibly at the two extremes of the interval: we must exclude intervals no longer than $(3N)2^{-n_1}$ on the left and an interval of length $(2N)2^{-n_1}$ on the right. The measure of the exceptional points $\{x : d^\tau_{n_1}(x) \neq d_{n_1}(x)\}$ is the sum of contributions from each complementary interval. Each such contribution is at most $5N \cdot 2^{-n_1}$, and as such, the total measure of the exceptional set is less than at most $5N$ times the measure of $\{x : \tau(x) = n_1\}$. We use these exceptional intervals to define a modification ρ of the stopping time τ. Each component of $\{x : \tau(x) = n_1\}$ is expanded to include an interval of length $(3N)2^{-n_1}$ on the right (the left extreme of the complementary interval) and an interval of length $(2N) \cdot 2^{-n_1}$ on the left (the right extreme of the complementary interval). Since each complementary interval is of length at least $(8N)2^{-n_1}$, we have diminished, but not eliminated, any complementary interval. The expanded component becomes a component of $\{x : \rho(x) = n_1\}$.

Step 10. We repeat this procedure for each $n > n_1$. On each component of the set $\{x : \tau(x) > n\}$, the analysis made in Steps 7, 8, and 9 may be applied. The functions $\tilde{f}^\tau_n(x)$ and $d^\tau_n(x)$ satisfy the same estimates. That is, $|\tilde{f}^\tau_n(x)| = O(\lambda)$ and $|d^\tau_n(x)| = O(\lambda)$.

Now let us write

$$f^\tau_n(x) = f_0(x) + \sum_{k=1}^{n} d^\tau_k(x).$$

The functions $d^\tau_n(x) = d_n(x)$ except on a set of measure comparable to the measure of the set $\{x : \tau(x) \leq n\}$. We prove this by induction, as follows: For $n = n_1$, we have shown that $d^\tau_{n_1}(x) = d_{n_1}(x)$ except on the set $\{x : \rho(x) = n_1\}$. Now fix a complementary interval of the set $\{x : \tau(x) = n_1\}$ (which contains a complementary interval of $\{x : \rho(x) = n_1\}$). There are two cases to consider: (i) The set $\{x : \tau(x) = n_1 + 1\}$ does not intersect the fixed complementary interval. (ii) The set $\{x : \tau(x) = n_1 + 1\}$ does intersect, creating a smaller complementary interval (or intervals).

In case (i), the difference $d^\tau_{n_1+1}(x) = d_{n_1+1}(x)$ except on intervals of length $(3N)2^{-(n_1+1)}$ and $(2N)2^{-(n_1+1)}$ on the left and right extremes of the complementary interval. However, these exceptional intervals are already contained in the complement of $\{x : \rho(x) = n_1\}$, so that no adjustment is needed in this case. The same is true for any $n > n_1$ as long as the set $\{x : \tau(x) = n\}$ does not intersect the complementary interval of $\{x : \tau(x) = n_1\}$. In other words, $d^\tau_n(x) = d_n(x)$ except on the set $\{x : \rho(x) = n_1\} \cap \{x : \tau(x) > n\}$.

In case (ii), the set $\{x : \tau(x) = n_1 + 1\}$ intersects the complementary interval of $\{x : \tau(x) = n_1\}$, creating a complementary interval (or intervals) to the set $\{x : \tau(x) \leq n_1 + 1\}$. In this case, we expand the components of $\{x : \tau(x) = n_1 + 1\}$ on the left and right by $(2N)2^{-(n_1+1)}$ and $(3N)2^{-(n_1+1)}$, respectively, if these components disconnect the complementary interval of $\{x : \tau(x) = n_1\}$. If the component of $\{x : \tau(x) = n_1 + 1\}$ does not disconnect (e.g., it falls, for example, at the right end of the complementary interval) we expand unilaterally (on the left, in the example).

This expansion process creates a new set of exceptional points $\{x : \rho(x) = n_1 + 1\}$, whose measure is comparable to the measure of $\{x : \tau(x) = n_1 + 1\}$.

We continue this process for all $n > n_1$, and so define the stopping time ρ.

Step 11. Now we stop the sequence f_n^τ, $n \geq 0$, using the stopping time ρ in the same way as in Step 8. At the level n, the complementary intervals of $\{x : \rho(x) \leq n\}$ are isolated and for each interval we define

$$f_n^\rho(x) = \sum_{s=\ell}^{r} b_{2s}\psi(2^n x - 2s).$$

As in Step 8 we take the shortest sum, so that $f_n^\rho(x) = f_n(x)$ on the complementary interval. The coefficients b_{2s} are, of course, the same as those in the expansion of d_n^τ. The new differences are denoted d_n^ρ.

We may now estimate the $\sup_n \|f_n^\rho\|_2$ as follows: On the set $\{x : \rho(x) = \infty\} \subset \{x : \tau(x) = \infty\}$,

$$(f^\rho)^*(x) = (\tilde{f}^\tau)^*(x) = O(\lambda),$$

as indicated in Step 10.

Therefore it remains for us to estimate $(f^\rho)^*(x)$ on the set $\{x : \rho(x) < \infty\}$. This set is a union of disjoint sets $\{x : \rho(x) = n\}$, $n = n_1, n_1 + 1, \ldots$; we estimate $(f^\rho)^*$ on each of these sets. Each set $\{x : \rho(x) = n\}$ is itself a union of components I_n^j, each of which is at least of length 2^{-n}. (Recall that a component of $\{x : \rho(x) = n\}$ contains a component of $\{x : \tau(x) = n\}$ of length less than $N \cdot 2^{-n}$ that is either contiguous with a component of $\{x : \tau(x) \leq n - 1\}$, of length greater than $8 \cdot N \cdot 2^{-(n-1)}$, or is an "endpoint interval," one that contains zero or one. Since we are restricting attention to $[0, 1]$, the endpoint components may be considered to have infinite length. Otherwise, any component of $\{x : \tau(x) = n\}$ (and $\{x : \rho(x) = n\}$) has length greater than or equal to $8 \cdot N \cdot 2^{-n}$.) By construction,

$$\max_{k<n} |f_k^\rho(x)| = O(\lambda),$$

so that it is necessary to estimate the magnitude of sums

$$\left|\sum_{k=n}^{m} d_k^\rho(x)\right|, \qquad m = n, n+1, \ldots,$$

on a fixed component of $\{x : \rho(x) = n\}$. Each of the differences

$$\left|d_{n+k}^\rho\right| = O(\lambda), \qquad k = 0, 1, \ldots.$$

Now consider that part of the support of d_{n+k}^ρ that lies in the component I_n^j of $\{x : \rho(x) = n\}$ under examination.

(a) If $k > 4\log_2 N$ (so that $N \cdot 2^{-(n+k)} < 2^{-n}$), then that part of the support of d_{n+k}^ρ that intersects the component interval I_n^j, is contained in at most two disjoint intervals, each of length $N \cdot 2^{-(n+k)}$, lying at either end of the component interval I_n^j. Thus, the supports of d_{n+k}^ρ form a decreasing sequence of sets within each component of $\{x : \rho(x) = n\}$. Therefore, we may majorize

$$\sum_{k>\log_2 4N} \left|d_{n+k}^\rho(x)\right|$$

in terms of the relative distance function $\Delta_n^j(x)$, defined on each component I_n^j of $\{x : \rho(x) = n\}$: Let $|I_n^j|$ be the length of I_n^j and

$$\frac{\Delta_n^j(x) = \text{distance } (x, \text{ complement of } I_n^j)}{|I_n^j|.}$$

The above considerations lead us to the estimate

$$\sum_{k>\log_2 4N} |d_{n+k}^{\tau}(x)| = O(\lambda)|\log_2 \Delta_n^j(x)|$$

for $x \in I_n^j$.

(b) If $k \leq \log_2 4N$, the best estimate is simply

$$|d_{n+k}^{\rho}(x)| = O(\lambda).$$

With these estimates, we may estimate $\|(f^\rho)^*\|_2$ as follows: On each component I_n^j,

$$\begin{aligned}|(f^\rho)^*(x)|^2 &\leq O(\lambda^2) + \left(\sum_{k\geq n} |d_k^\rho(x)|\right)^2 \\ &\leq O(\lambda^2)\left[1 + \log_2^2(N) + \log_2^2 \Delta_n^j(x)\right].\end{aligned}$$

Therefore

$$\int_{I_n^j} |(f^\rho)^*(x)|^2\, dx = O(\lambda^2)|I_n^j| + O(\lambda^2)\int_{I_n^j} \log_2^2 \Delta_n^j(x) dx.$$

The last integral is estimated by the quantity

$$\left(\int_0^1 \log_2^2 |x|^{-1} dx\right)|I_n^j|.$$

If we sum these estimates over j and n, we obtain

$$\begin{aligned}\int |(f^\rho)^*(x)|^2 dx &= \int_{\{\rho(x)<\infty\}} + \int_{\{\rho(x)=\infty\}} |(f^\rho)^*(x)|^2 \\ &= O(\lambda^2).\end{aligned}$$

Step 12. The stopped sequence $f_n^\rho(x)$, $n \geq 0$, agrees with the original sequence $f_n(x)$, $n \geq 0$ except for points in a set of measure comparable to the measure of $\{x : \rho(x) < \infty\}$. This set has small measure (less than $O(\epsilon)$). Furthermore, the estimate given in the previous step shows that $\|(f^\rho)^*\|_2 = O(\lambda)$, and f_n^ρ, $n \geq 0$, satisfies the property $P_n f_{n+1}^\rho = f_n^\rho$. Therefore

$$\begin{aligned}\|S(f^\rho)\|_2 &= \sup_n \|f_n^\rho\|_2 \\ &\leq \|(f^\rho)^*\|_2 \\ &= O(\lambda).\end{aligned}$$

Since $S(f^\tau)(x) = S(f)(x)$ on the set where $f_n^\rho(x) \equiv f_n(x)$, $n = 0, 1, \ldots$, we may conclude that meas$(B\setminus A) = O(\epsilon)$. Furthermore, it is known that L^2-bounded wavelet expansions converge a.e. in this context (see [7]), so that meas$(B \setminus A) = 0$.

Step 13. The entire procedure may be repeated using B as the initial set. The conclusion of this argument is meas$(A \setminus B) = 0$, and the proof of Theorem B is now complete. □

Acknowledgment. We would like to thank the Volkswagen Foundation for support during July 1996 at the Mathematics Research Institute, Oberwohlfach, Germany.

REFERENCES

[1] D. L. Burkholder and R. F. Gundy, *Extrapolation and interpretation of quasi-linear operators on martingales*, Acta Math., 124 (1970), pp. 250–304.

[2] C. K. Chui, *Introduction to Wavelets*, Academic Press, New York, 1992.

[3] R. F. Gundy, *Martingale theory and the pointwise convergence of certain orthogonal series*, Trans. Amer. Math. Soc., 124 (1966), pp. 228–248.

[4] R. F. Gundy, *The martingale version of a theorem of Marcinkiewicz and Zygmund*, Ann. Math. Statist., 38 (1967), pp. 725–734.

[5] F. G. Arutyunyan, *On the series in the Haar system*, Akad. Nauk. Armenii Dokl., 42 (1966), pp. 134–140.

[6] E. Hernández and G. Weiss, *A First Course on Wavelets*, CRC Press, Boca Raton, FL, 1996.

[7] S. E. Kelly, M. A. Kon, and L. A. Raphael, *Pointwise convergence of wavelet expansions*, Bull. Amer. Math. Soc. (N.S.), 30 (1994), pp. 87–94.

[8] J. Marcinkiewicz and A. Zygmund, *Sur les fonctions indépendantes,* Fund. Math., 29 (1937), pp. 60–90.

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 574–580

STABILITY PROPERTIES FOR A COMPACTLY SUPPORTED PRESCALE FUNCTION[*]

V. DOBRIĆ[†], R. F. GUNDY[‡], AND P. HITCZENKO[§]

Abstract. We show that if ϕ is a continuous, minimally supported prescale function, then its translates are linearly independent on any set of positive measure in the unit interval. This generalizes results of Y. Meyer and P. G. Lemarié.

This result implies that a stability condition, introduced by Gundy and Kazarian for the study of local convergence of spline wavelet expansions, is satisfied for all expansions arising from multiresolution analyses generated by such prescale functions ϕ.

Key words. wavelets, linear independence, local convergence

AMS subject classification. 42C15

PII. S003614109732746X

1. Introduction.

In [7], P. G. Lemarié proved that if a multiresolution analysis contains a compactly supported function, then it contains a minimal (pre)scale function. More specifically, there exists a function ϕ of compact support such that

(1) the integer translates, $\phi(\cdot - k),\ k \in \mathbf{Z}$, are a Riesz basis for the space V_0;

(2) every function in V_0 that is compactly supported may be written as a finite linear combination of translates of ϕ.

(Throughout this paper, we will use the term "scale function" to mean the above, although some authors refer to this as "prescale function.") The most basic examples of minimal scale functions of this type are the B-splines and the compactly supported scale functions constructed by I. Daubechies [1]. An important property of these minimal-scale functions was first proved by Y. Meyer [10] for Daubechies' functions and subsequently stated by Lemarié [7] in the general case: *The translates of ϕ, restricted to the unit interval, form a linearly independent set.*

The purpose of this paper is to prove the following stronger version of the above for a minimal scale function ϕ that is continuous on the unit interval: the translates of ϕ are linearly independent over any subset of positive measure contained in the unit interval. The stronger version is of interest because it may be used to obtain a local convergence theorem for multiresolution analyses with continuous minimal scale functions. The first version of this type of local convergence theorem was proved by Gundy and Kazarian [4] for a class of wavelet expansions that includes the spline wavelets. Their theorem assumed a stability condition (condition (M-Z) of [4]). It turns out that this stability condition is, in fact, a property of all multiresolution analyses with continuous minimal scale functions, as a consequence of the above strong linear independence of these functions.

[*] Received by the editors September 18, 1997; accepted for publication (in revised form) May 19, 1999; published electronically February 24, 2000.

http://www.siam.org/journals/sima/31-3/32746.html

[†] Department of Mathematics, Lehigh University, Bethlehem, PA 18015 (vd00@lehigh.edu).

[‡] Department of Statistics, Rutgers University, New Brunswick, NJ 08903 (gundy@rci.rutgers.edu).

[§] Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (pawel@math.ncsu.edu). This author's research was carried out while he was visiting Rutgers University and was partially supported by the National Science Foundation.

2. Notation. We suppose that a multiresolution analysis is given. That is, we have a sequence of subspaces of $L^2(\mathbf{R})$, V_j, $j \in \mathbf{Z}$, such that $V_j \subset V_{j+1}$, and $f(\cdot) \in V_j$ iff $f(2^{-j}\cdot) \in V_0$. Furthermore, we are given a function $\phi \in V_0$ such that the integer translates $\phi(\cdot - k)$ form a Riesz basis for V_0: any function $f(\cdot) \in V_0$ has a representation

$$f(x) = \sum a_k \phi(x-k)$$

with

$$\sum a_k^2 \cong \|f\|_2^2, \quad \text{i.e.,} \quad c\sum a_k^2 \leq \|f\|_2^2 \leq C \sum a_k^2, \quad 0 < c < C.$$

If ϕ has a nonzero integral, then it follows that the increasing sequence of subspaces exhausts $L^2(\mathbf{R})$ (see [5, Chapter 2]). Let P_j be the orthogonal projection operator from L^2 onto V_j.

Now we impose another restriction on the multiresolution analysis. We require that the space V_0 contains a nontrivial continuous function that is compactly supported. With this additional assumption, the techniques of Lemarié [7] may be used to show that there exists a minimally supported, real-valued, continuous function $\phi \in V_0$ such that every compactly supported function in V_0 admits a representation as a finite linear combination of integer translates of ϕ. If we agree to normalize ϕ by setting its integral equal to one, then ϕ is unique, up to integer translates. (Our class of multiresolution analyses does include the spline wavelets, the compactly supported Daubechies wavelets, and those obtained from these classes by integration, as indicated in Lemarié [7].)

3. Linear independence of translates. In this section, we state the theorem on linear independence.

THEOREM 1. *Let ϕ be a continuous, minimal (pre)scale function supported on the interval* $[0, N]$. *Then the translates* $\phi(\cdot + k)$, $k = 0, \ldots, N-1$, *are linearly independent over any set of positive measure of the unit interval.*

Remarks. As we noted above, this line of investigation was initiated by Y. Meyer [10] and pursued by P. G. Lemarié in [7]. These authors treated the case where the "set of positive measure" was the entire unit interval. Lemarié and Malgouyres [8] gave another simplified proof that showed that the translates were linearly independent on any subinterval of the unit interval. Finally, Lemarié [7] showed that this property characterizes minimal scale functions. Those authors made no continuity assumptions.

Proof. We give a proof by contradiction as follows: If the translates are linearly dependent over a set of positive measure, we show that they are dependent over a set of measure one in the unit interval. Since the function ϕ is continuous, this means that the translates of ϕ are dependent over the unit interval itself, thus contradicting the theorem of Meyer.

Throughout the proof, we will use matrices $\mathbf{P}_0$ and $\mathbf{P}_1$. To define these matrices, let us write the dilation equation for ϕ as

$$\phi(x/2) = \sum_{k=0}^{N} p_k \phi(x-k) \qquad \text{with } p_0, p_N \neq 0.$$

With this notation, let us define the $(N-1)\times(N-1)$ matrix $\mathbf{P}$ whose first row consists of the vector of odd numbered coefficients, p_{2k+1}, followed by the appropriate number of zeros to give the vector $N-1$ components. The second row of $\mathbf{P}$ is defined in the same way, using the even numbered coefficients, p_{2k}, followed by the appropriate number of zeros. Third and fourth rows are obtained from the first two rows by a cyclic permutation of the indices: each entry is shifted to the right, with the final entry, a zero, moving to first position. This procedure is continued until $N-1$ rows are obtained. (Thus if $N=2k$, the second row will contain the $k+1$ entries $p_0, p_2, \ldots, p_{2k}$ followed by $k-2$ zeros. The last row will contain $k-1$ zeros followed by the k coefficients $p_1, p_3, \ldots, p_{2k-1}$. If $N=2k+1$, then the last row of the matrix consists of $k-1$ zeros, followed by the $k+1$ entries $p_0, p_2, \ldots, p_{2k}$.) Now define the two $N \times N$ matrices

$$\mathbf{P}_0 = \begin{pmatrix} p_0 & \mathbf{p}_t \\ 0 & \mathbf{P} \end{pmatrix} \qquad \text{and} \qquad \mathbf{P}_1 = \begin{pmatrix} \mathbf{P} & 0 \\ \mathbf{p}_b & p_N \end{pmatrix},$$

where $\mathbf{p}_t$ is the $N-1$ vector consisting of the even numbered p_k, starting with p_2, followed by the appropriate number of zeros; $\mathbf{p}_b$ consists of zeros followed by the coefficients p_k where k has the same parity as N, where the final entry of the vector $\mathbf{p}_b$ is the coefficient p_{N-2}.

The roles of $\mathbf{P}_0$ and $\mathbf{P}_1$ are as follows. Consider a general linear combination of translates $\sum c_k \phi(x+k)$. If we take into account fact that ϕ is supported on $[0,N]$ and restrict attention to $x \in [0,1]$, this sum is, in fact, finite and may be expressed as $\sum_{k=0}^{N-1} c_k \phi(x+k)$. If we apply the dilation equation to express each $\phi(\cdot + k)$ in terms of a sum of translates of $\phi(2\cdot)$, the resulting double sum is a certain linear combination of translates of $\phi(2x)$ and $\phi(2x-1)$, depending on whether x is in $[0, \frac{1}{2}]$ or in $[\frac{1}{2}, 1]$. The coefficients of this linear combination are given by the matrices $\mathbf{P}_0$ or $\mathbf{P}_1$, acting on the vector $\mathbf{c} = (c_0, \ldots, c_{N-1})$. (These matrices are implicit in the reconstruction-decomposition schemes in the wavelet literature and appear explicitly in the 3×3 case in Daubechies [2, section 7.2].) Let $\mathbf{\Phi}(x) = \big(\phi(x), \phi(x+1), \ldots, \phi(x+N-1)\big)^t$ for $x \in [0,1]$, and let $\epsilon_k(x)$, $k = 1, 2, \ldots$, be the digits in the binary expansion of x. That is, $x = \sum \epsilon_k / 2^k$, with $\epsilon_k = 0$ or 1, $k \geq 1$. Let T be the plus-one shift on the ϵ-sequence: $T : (\epsilon_1, \epsilon_2, \ldots) \to (\epsilon_2, \epsilon_3, \ldots)$. We write $Tx = \sum_{k=1}^{\infty} \epsilon_{k+1}/2^k$. We summarize the above in the following lemma.

Lemma 1. *For* $\mathbf{c} = (c_0, c_1, \ldots, c_{N-1})$ *we have*

$$\mathbf{c} \circ \mathbf{\Phi}(x) = (\mathbf{P}_{\epsilon_1} \mathbf{c}^t) \circ \mathbf{\Phi}(Tx).$$

More generally, for any m,

$$\mathbf{c} \circ \mathbf{\Phi}(x) = (\mathbf{P}_{\epsilon_1} \cdots \mathbf{P}_{\epsilon_m} \mathbf{c}^t) \circ \mathbf{\Phi}(T^m x).$$

Proof. Recall that the support of $\phi(\cdot + m)$ is the interval $[-m, -m+N]$. For $x \in [0,1]$,

$$\begin{aligned} \mathbf{c} \circ \mathbf{\Phi}(x) &= \sum_{k=0}^{N-1} c_k \phi(x+k) = \sum_k c_k \bigg(\sum_j p_j \phi\big(2(x+k)-j\big) \bigg) \\ &= \sum_m \bigg(\sum_k c_k p_{2k-m} \bigg) \phi(2x+m) = \sum_m \bigg(\sum_k c_k p_{2k-m} \bigg) \phi(Tx + \epsilon_1 + m). \end{aligned}$$

The inner sum is taken over all k with the provision that $p_{2k-m} = 0$ if $2k-m$ is not one of the integers $0, 1, \ldots, N$. The outer sum with index m changes according to whether $0 < 2x \leq 1$ or $1 < 2x \leq 2$, due to the support condition mentioned above. In the first case, when $\epsilon_1 = 0$, we have $0 \leq m \leq N-1$; in the second case, when $\epsilon_1 = 1$,

$0 \le m+1 \le N-1$. Thus, the transformation takes two forms, with matrices $\mathbf{P}_0$ and $\mathbf{P}_1$. The rest of Lemma 1 is easily proved by induction.

Fix $\mathbf{c} = (c_0, \ldots, c_{N-1})$ and let $K_{\mathbf{c}} = \{x : \mathbf{c} \circ \mathbf{\Phi}(x) = 0\}$. The continuity of $\mathbf{\Phi}(x)$ implies that $K_{\mathbf{c}}$ is closed. From now on, we assume that $\mathbf{c} \neq 0$ and that $K_{\mathbf{c}}$ has positive measure in $[0,1]$; we seek to contradict Meyer–Lemarié theorem [10], [7]. There are two cases to consider. □

Case 1. There exists a finite sequence $\mathbf{P}_{\epsilon_k}$, $k = 1, \ldots, m$, such that $\mathbf{P}_{\epsilon_1} \cdots \mathbf{P}_{\epsilon_m} \mathbf{c}^t = 0$. If this is the case, we have our contradiction since $\mathbf{c} \circ \mathbf{\Phi} \equiv 0$ on the dyadic interval

$$\{x : \epsilon_1(x) = \epsilon_1, \ldots, \epsilon_m(x) = \epsilon_m\}.$$

Case 2. The vector $\mathbf{c}$ is such that $\mathbf{P}_{\epsilon_1} \cdots \mathbf{P}_{\epsilon_m} \mathbf{c}^t \neq 0$ for every finite sequence $\epsilon_1, \ldots, \epsilon_m$. In this case, we say that $\mathbf{c}$ is a "never zero" vector.

Lemma 2. *Let* $\mathbf{c}$ *be a never zero vector. Then, for every* η, $0 < \eta < 1$, *there exists a never zero vector* $\mathbf{b}$ *such that* $m(K_{\mathbf{b}}) > 1 - \eta$.

Proof. Since $K_{\mathbf{c}}$ has positive measure, we can find a dyadic interval $I_j = \{x : \epsilon_1(x) = \epsilon_1, \ldots, \epsilon_j(x) = \epsilon_j\}$ such that $m(K_{\mathbf{c}} \cap I_j)/2^{-j} > 1 - \eta$. This is a consequence of the fact that every Lebesgue measurable set may be approximated by a finite union of dyadic intervals. Set $\mathbf{b} = \mathbf{P}_{\epsilon_1} \cdots \mathbf{P}_{\epsilon_j} \mathbf{c}^t$. Then $T^j(K_{\mathbf{c}} \cap I_j) \subseteq K_{\mathbf{b}}$ by Lemma 1. The set $K_{\mathbf{b}}$ has measure greater than $1 - \eta$ since $m\big(\{x : x = T^j y \text{ for some } y \in I_j\}\big) = 1$. This proves Lemma 2. □

Now set

$$\mathcal{A}_{\mathbf{c}} = \left\{ \mathbf{a} \in \mathbb{R}^N : \mathbf{a} = \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \text{ for some } \mathbf{b} = \mathbf{P}_{\epsilon_1} \cdots \mathbf{P}_{\epsilon_j} \mathbf{c}, \quad j \in \mathbf{Z} \right\}.$$

By Lemma 2, we have

$$\sup_{\mathbf{a} \in \mathcal{A}_{\mathbf{c}}} m(K_{\mathbf{a}}) = 1.$$

Now we claim that there is an $\mathbf{a} \in \mathbf{R}^N$ with $\|\mathbf{a}\|_2 = 1$ such that $m(K_{\mathbf{a}}) = 1$. To this end, we topologize the class $\mathcal{K}$ of all nonempty compact subsets of $[0,1]$ with the Hausdorff metric ρ:

$$\rho(A,B) = \sup_{x \in [0,1]} |d(x,A) - d(x,B)|,$$

where $d(x,D) = \inf\{|x-y| : y \in D\}$. This metric is equivalent to

$$\sigma(A,B) = \inf\{\epsilon > 0, A \subset V_\epsilon(B) \text{ and } B \subset V_\epsilon(A)\},$$

where $V_\epsilon(D) = \{z \in [0,1] : d(z,D) < \epsilon\}$. It is known that $(\mathcal{K}, \rho)$ is a compact metric space. A complete discussion of these facts may be found in Kornum [6, section 6.2].

Let $\{K_{\mathbf{a}_n}\}$ be a sequence of sets in $\mathcal{K}$ such that $m(K_{\mathbf{a}_n})$ tends to one. From this sequence we may extract a convergent subsequence $\{K_{\mathbf{a}_{n_k}}\}$ with limit K. Since $\|\mathbf{a}_{n_k}\|_2 = 1$, we may extract a convergent subsequence of $\mathbf{a}_{n_k}$, call it $\{\mathbf{a}_n\}$, so that finally, we obtain sequences $K_{\mathbf{a}_n} \to K$ and $\mathbf{a}_n \to \mathbf{a}$. Since $E \to m(E)$ is an uppersemicontinuous function on $(\mathcal{K}, \rho)$, that is if $E_n \to E$ then $\overline{\lim} m(E_n) \le m(E)$, it follows that $m(K) = 1$. Second, we claim that $K \subset K_{\mathbf{a}}$. Since $K_{\mathbf{a}_n} \to K$, we have

$$\sup_{y \in [0,1]} |d(y,K) - d(y, K_{\mathbf{a}_n})| \to 0.$$

Therefore, if $x \in K$, $d(x, K_{\mathbf{a}_n}) \to 0$. Choose $x_n \in K_{\mathbf{a}_n}$ so that $|x - x_n| = d(x, K_{\mathbf{a}_n})$. Then, by the Cauchy–Schwarz inequality and the definition of $K_{\mathbf{a}}$,

$$\begin{aligned} |\mathbf{a} \circ \Phi(x)| &\le |(\mathbf{a} - \mathbf{a}_n) \circ \mathbf{\Phi}(x)| + |\mathbf{a}_n \circ (\mathbf{\Phi}(x) - \mathbf{\Phi}(x_n))| + |\mathbf{a}_n \circ \mathbf{\Phi}(x_n)| \\ &\le \|\mathbf{a} - \mathbf{a}_n\|_2 \cdot \|\mathbf{\Phi}(x)\|_2 + \|\mathbf{\Phi}(x) - \mathbf{\Phi}(x_n)\|_2. \end{aligned}$$

Since both terms on the right tend to zero, we have the inclusion $K \subset K_{\mathbf{a}}$. Therefore, $m(K_{\mathbf{a}}) = 1$. □

4. Local convergence of wavelet expansions. In [3], the following local convergence theorem is proved for Haar series, using martingale methods.

Theorem A. *Let $\{f_j\}$ be a sequence of functions such that*

(a) *$f_j \in V_j$ where $\{V_j\}$ is the Haar multiresolution analysis;*

(b) *$P_j(f_{j+1})(x) = f_j(x)$ for $j \ge 0$.*

Set $f(x) = (f_0(x), f_1(x), \dots)$ and let $S(f)(x) = \left(\sum (f_{j+1}(x) - f_j(x))^2 + f_0^2(x)\right)^{1/2}$; $f^*(x) = \sup_j |f_j(x)|$. Then, the following sets are equivalent almost everywhere (a.e.):

(a) *$\{x : \lim_{j\to\infty} f_j(x)$ exists and is finite$\}$;*

(b) *$\{x : S(f)(x) < +\infty\}$;*

(c) *$\{x : f^*(x) < \infty\}$.*

Gundy and Kazarian [4] extended this local convergence theorem to the class of multiresolution analyses arising from the basic splines. In fact, the proof did not appear to use properties specific to the spline family. The basic stability condition essential to the proof is a two-norm condition, reminiscent of a condition first proposed by Marcinkiewicz and Zygmund [9] in their study of series of independent random variables. This condition, called condition (M-Z), is as follows: Let ϕ be a compactly supported scale function, supported on $[0, N]$. We suppose that, for every δ, $0 < \delta < 1$, there exist constants B_δ and C_δ such that for every measurable subset $E \subset [0, 1]$ of measure greater than δ and any sequence a_k; $k = 0, 1, \dots, N-1$, we have

$$\begin{aligned} C_\delta \sum_{k=0}^{N-1} |a_k| &\le \sup_{x\in E} \left| \sum_{k=0}^{N-1} a_k \phi(x+k) \right| \\ &\le B_\delta \sum_{k=0}^{N-1} |a_k|. \end{aligned}$$

The constants B_δ, C_δ depend only on ϕ and δ. The condition holds for the class of B-spline scale functions, as pointed out in [4]. However, the scope of the condition was not known and was left as an open problem in [4]. The following proposition we label as a corollary of Theorem 1.

Corollary 1. *Let $\{V_j\}$ be a multiresolution analysis such that V_0 contains a continuous function of compact support. Then the minimal scale function ϕ satisfies condition* (M-Z).

Before proving the corollary, we state the following theorem, in which we use the definitions in Theorem A.

Theorem 2 (Theorem B of [4]). *Let $\{V_j\}$ be a multiresolution analysis that contains a continuous function of compact support. Then the following sets are equivalent a.e.:*

(a) *$\{x : \lim_{j\to\infty} f_j(x)$ exists and is finite$\}$;*

(b) *$\{x : S(f)(x) < +\infty\}$;*

(c) *$\{x : f^*(x) < \infty\}$.*

Proof of Corollary 1. Notice that B_δ may be taken to be $\|\phi\|_\infty$. Since ϕ is continuous on $[0,1]$, the issue is to show the existence of C_δ that is uniform over all sets $E \subset [0,1]$ of measure greater than δ. First, observe that, since the translates of ϕ are linearly independent over E (Theorem 1), there is a constant $C(E) > 0$, such that

$$C(E) \sum_{k=0}^{N-1} |a_k| \leq \sup_{x\in E} \left| \sum_{k=0}^{N-1} a_k \phi(x+k) \right|.$$

This follows from the fact that the right-hand side defines a norm on $\mathbf{R}^N$: the linear independence of the translates of ϕ on the set E guarantees that the right-hand side is strictly positive on $\mathbf{R}^N \setminus \{0\}$. Since the left-hand side is also a norm, the existence of a constant is assured by the equivalence of norms on finite dimensional spaces. Now we must show that

$$\inf\{C(E) : m(E) \geq \delta\} > 0.$$

It is enough to show this for closed sets. To this end, we show that $C : (\mathcal{K}, \rho) \to \mathbf{R}$ defined by

$$C(E) = \inf_{\mathbf{a}\neq 0} \sup_{x\in E} \frac{|\mathbf{a} \circ \mathbf{\Phi}(x)|}{\|\mathbf{a}\|_1}$$

is a continuous function. Let $\epsilon > 0$ be given, and let $\{A_n\}$ be a sequence of sets converging to A in $\mathcal{K}$. The function $\mathbf{\Phi}$ is uniformly continuous on $[0,1]$; that is,

$$\|\mathbf{\Phi}(x) - \mathbf{\Phi}(y)\|_2 \leq \epsilon$$

whenever $|x-y| \leq \eta(\epsilon)$. Let n_0 be an integer such that

$$A_n \subset V_\eta(A) \qquad \text{and} \qquad A \subset V_\eta(A_n)$$

for all $n \geq n_0$. Notice that

$$C(A) \leq C(\overline{V_\eta(A_n)}),$$

and that, by continuity, there exists $x = x(n, \eta, \mathbf{a}) \in \overline{V_\eta(A_n)}$ such that

$$C(\overline{V_\eta(A_n)}) = \inf_{\mathbf{a}\neq 0} \frac{|\mathbf{a} \circ \mathbf{\Phi}(x)|}{\|\mathbf{a}\|_1}.$$

Choose $y = y(n, \eta, \mathbf{a}) \in A_n$ so that $|x-y| \leq \eta(\epsilon)$. Then

$$\begin{aligned} C(\overline{V_\eta(A_n)}) &\leq \inf_{\mathbf{a}\neq 0} \left(\frac{\|\mathbf{a}\|_2}{\|\mathbf{a}\|_1} \|\mathbf{\Phi}(x) - \mathbf{\Phi}(y)\|_2 + \frac{|\mathbf{a} \circ \mathbf{\Phi}(y)|}{\|\mathbf{a}\|_1} \right) \\ &\leq \inf_{\mathbf{a}\neq 0} \left(\frac{\|\mathbf{a}\|_2}{\|\mathbf{a}\|_1} \epsilon + \sup_{y\in A_n} \frac{|\mathbf{a} \circ \mathbf{\Phi}(y)|}{\|\mathbf{a}\|_1} \right) \\ &\leq \inf_{\mathbf{a}\neq 0} \left(\epsilon + \sup_{y\in A_n} \frac{|\mathbf{a} \circ \mathbf{\Phi}(y)|}{\|\mathbf{a}\|_1} \right) \\ &= \epsilon + C(A_n). \end{aligned}$$

If we reverse the roles of A_n and A in the above argument, we see that $C(A_n) \leq C(A) + \epsilon$. Thus, C is continuous on $\mathcal{K}$.

The collection $\{E \in \mathcal{K} : m(E) \geq \delta\}$ is a closed set in $\mathcal{K}$ since $m(\cdot)$ is uppersemicontinuous on $(\mathcal{K}, \rho)$. Therefore, by continuity of C, there exists an $E_0 \in \mathcal{K}$, $m(E_0) \geq \delta$ such that $C(E_0) \leq C(E)$ for all $E \in \mathcal{K}$ with $m(E) \geq \delta$. Now, by the compactness of the unit sphere in ℓ_1^N, there exists an $\mathbf{a}$ such that

$$C(E_0) = \sup_{x \in E_0} \frac{|\mathbf{a} \circ \mathbf{\Phi}(x)|}{\|\mathbf{a}\|_1}.$$

By Theorem 1, $C(E_0) > 0$. □

5. Concluding remarks. The quadratic variation functional $S(f)$ of Theorem 2 is invariant under changes of scale functions ϕ for V_0 and ψ for V_1: $S(f)$ is defined from the sequence of projections $\{P_j\}$ without specific reference to the choice of scale function. However, $S(f)$ is an "incomplete" square function in the sense that if the prewavelet family $\{\psi(2^j \cdot -k)\}$ is orthogonalized in k to obtain a family $\{\tilde{\psi}(2^j \cdot -k)\}$ that is orthonormal in both variables $j, k \in \mathbf{Z}$, then one could consider a quadratic variation functional $\mathbf{S}(f)(x) = \left(\sum_{j,k}(a_{j,k}\tilde{\psi}(2^j x - k))^2\right)^{1/2}$. If we have a multiresolution analysis that admits a compactly supported, continuous orthonormal family $\{\tilde{\psi}(2^j \cdot -k)\}$, then one can show that $S(f)$ and $\mathbf{S}(f)$ are finite on the same set, up to a set of measure zero. The proof of this fact follows the same lines as the proof in [4]. Since the details are given there, we will not repeat them here.

Acknowledgments. P. Hitczenko would like to thank R. F. Gundy, R. Wilson, and the Departments of Mathematics and Statistics at Rutgers University for their hospitality.

REFERENCES

[1] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[2] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, PA, 1992.

[3] R. F. Gundy, *Martingale theory and the pointwise convergence of certain orthogonal series*, Trans. Amer. Math. Soc., 124 (1966), pp. 228–248.

[4] R. F. Gundy and K. Kazarian, *Stopping times and local convergence for spline wavelet expansions*, SIAM J. Math. Anal, 31 (2000), pp. 561–573.

[5] E. Hernández and G. Weiss, *A First Course on Wavelets*, CRC Press, Boca Raton, FL, 1996.

[6] P. Kornum, *Construction of Borel Measures on Metric Spaces*, Matematisk Institut, Aarhus Universitet, Aarhus, Denmark, 1980.

[7] P. G. Lemarié, *Fonctions à support compact dans les analyses multi-résolutions*, Rev. Mat. Iberoamericana, 7 (1991), pp. 157–182.

[8] P. G. Lemarié and G. Malgouyres, *Support des fonctions de base dans une analyse multi-résolution*, C. R. Acad. Sci. Paris Sér. I Math., 313 (1991), pp. 377–380.

[9] J. Marcinkiewicz and A. Zygmund, *Sur les fonctions indépendentes*, Fund. Math., 29 (1937), pp. 60–90.

[10] Y. Meyer, *Ondelettes sur l'intervalle*, Rev. Mat. Iberoamericana, 7 (1991), pp. 115–133.

SIAM J. Math. Anal.
Vol. 31, No. 3, pp. 581–624

ASYMPTOTIC BEHAVIOR OF NONLINEAR ELLIPTIC SYSTEMS ON VARYING DOMAINS*

JUAN CASADO DIAZ† AND ADRIANA GARRONI‡

Abstract. We consider a monotone operator of the form $Au = -\mathrm{div}(a(x, Du))$, with $\Omega \subseteq \mathbf{R}^n$ and $a : \Omega \times \mathbf{M}^{M\times N} \to \mathbf{M}^{M\times N}$, acting on $W_0^{1,p}(\Omega, \mathbf{R}^M)$. For every sequence (Ω_h) of open subsets of Ω and for every $f \in W^{-1,p'}(\Omega, \mathbf{R}^M)$, $1/p+1/p' = 1$, we study the asymptotic behavior, as $h \to +\infty$, of the solutions $u_h \in W_0^1(\Omega_h, \mathbf{R}^M)$ of the systems $Au_h = f$ in $W^{-1,p'}(\Omega_h, \mathbf{R}^M)$, and we determine the general form of the limit problem.

Key words. homogenization, perforated domains, Dirichlet systems

AMS subject classifications. 35B27, 35B40

PII. S0036141097329627

1. Introduction.

In this paper we study the asymptotic behavior of the solutions of elliptic nonlinear systems, of M equations and N variables, on varying domains with Dirichlet boundary conditions. Namely, let Ω be a bounded open subset of $\mathbf{R}^N$ and let $1 < p < +\infty$. We regard A as a vector monotone operator defined from $W^{1,p}(\Omega, \mathbf{R}^M)$ to $W^{-1,p'}(\Omega, \mathbf{R}^M)$, mapping $u \in W^{1,p}(\Omega, \mathbf{R}^M)$ in $Au = -\mathrm{div}\big(a(x, Du)\big) \in W^{-1,p'}(\Omega, \mathbf{R}^M)$. The function $a\colon \Omega \times \mathbf{M}^{M\times N} \mapsto \mathbf{M}^{M\times N}$ is a Carathéodory function which satisfies the standard assumptions of strong monotonicity and Hölder continuity (see conditions (i)–(iv) in section 5). Given an arbitrary sequence of open subsets Ω_n of Ω and given an arbitrary $f \in W^{-1,p'}(\Omega, \mathbf{R}^M)$, we consider the solutions u_n of the following systems with Dirichlet boundary condition

$$u_n \in W_0^{1,p}(\Omega_n, \mathbf{R}^M)\,, \qquad Au_n = f \quad \text{in } \Omega_n. \tag{1.1}$$

We set $u_n = 0$ in $\Omega \setminus \Omega_n$ and regard the sequence (u_n) as a sequence in $W_0^{1,p}(\Omega, \mathbf{R}^M)$. Our results describe the asymptotic behavior of (u_n) as $n \to \infty$ and characterize the limit function as the solution of a variational "limit problem."

The main result of this paper is given by the following compactness theorem.

Theorem 1.1. *Let Ω_n be an arbitrary sequence of open subsets of Ω. Then there exist a subsequence of Ω_n, still denoted by Ω_n, a measure μ in the class $\mathcal{M}_0^p(\Omega)$ of positive Borel measures not charging set of p-capacity zero, and a vector function $F : \Omega \times \mathbf{R}^M \to \mathbf{R}^M$, such that for every $f \in W^{-1,p'}(\Omega, \mathbf{R}^M)$ the sequence (u_n) of solutions of problems* (1.1) *converges weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ to the solution u of the variational problem*

$$\begin{cases} u \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M)\,, \\ \displaystyle\int_\Omega a(x, Du)Dv\,dx + \int_\Omega F(x, u)v\,d\mu \;=\; \langle f, v\rangle \\ \forall v \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M). \end{cases} \tag{1.2}$$

*Received by the editors November 5, 1997; accepted for publication (in revised form) September 9, 1998; published electronically February 24, 2000.

http://www.siam.org/journals/sima/31-3/32962.html

†Departamento de ecuaciones differenciales, C. Tarfia Sin, 41012 Sevilla, Spain (jcasado@numer.us.es).

‡Dipartimento di Matematica, Università "La Sapienza," Piazzale A. Moro 2, 00185 Roma, Italy (garroni@mat.uniroma1.it).

By $L^p_\mu(\Omega,\mathbf{R}^M)$ we denote the standard L^p spaces with respect to the measure μ. Note that in this general case the usual "extra term" is given by $\int_\Omega F(x,u)v\,d\mu$.

The problem considered in the present paper has been studied, under various degree of generality, by many authors, with several approaches and in different frameworks. Most of the known results are given under assumptions involving the geometry or the capacity of the closed sets $\Omega\setminus\Omega_n$, which in general imply that the measure μ in the limit problem is a Radon measure (see, for instance, [20], [22], and [7] for the linear case, and [24], [25], [26], [27], [21], and [3], for monotone operators).

The class $\mathcal{M}^p_0(\Omega)$ described above also includes measures which take the value $+\infty$ on large families of sets; in this way, Dirichlet problems in subdomains of Ω can be written in the form (1.2) for a suitable choice of μ. Indeed, it is easy to see that, if E is a closed subset of Ω and the measure μ is defined by

$$\mu(B)=\begin{cases}0 & \text{if } C_p(B\cap E)=0\\ +\infty & \text{otherwise,}\end{cases} \tag{1.3}$$

for every Borel subset B of Ω, where C_p denotes the p-capacity, then problem (1.2) is equivalent to

$$u\in W^{1,p}_0(\Omega\setminus E,\mathbf{R}^M)\,,\qquad Au=f\quad\text{in }\Omega\setminus E.$$

In view of the latter remark, the compactness result above will be proved in a more general formulation (see Theorem 6.4) for a sequence of problems of the type

$$\begin{cases}u_n\in W^{1,p}_0(\Omega,\mathbf{R}^M)\cap L^p_{\mu_n}(\Omega,\mathbf{R}^M)\,,\\[4pt] \displaystyle\int_\Omega a(x,Du_n)Dv\,dx+\int_\Omega F_n(x,u_n)v\,d\mu_n\;=\;\langle f,v\rangle\\[4pt] \forall v\in W^{1,p}_0(\Omega,\mathbf{R}^M)\cap L^p_{\mu_n}(\Omega,\mathbf{R}^M)\,,\end{cases} \tag{1.4}$$

which for a suitable choice of (μ_n) in $\mathcal{M}^p_0(\Omega)$ reduce to (1.1), and includes also Schrödinger systems with positive oscillating potentials. A further motivation for the study of problem (1.4) is given by the recent applications to a relaxed formulation of some optimal design problems (see, for instance, [2]).

The compactness result in the setting of (1.4) was first proved for the scalar case $M=1$, using Γ-convergence techniques, in [13] and [14] when $p=2$ and A is a symmetric linear elliptic operator, and in [10] if A is $p-1$ homogeneous, under the assumption that it is the subdifferential of a convex functional. These results were generalized using Tartar's energy method in [11] for the general scalar linear case, and subsequently for the nonlinear case under an assumption of homogeneity of order $p-1$ for the operator A (see [15] and [16]). In these cases the extra term which appears in the limit problem is proved to be of the form $\int_\Omega |u|^{p-2}uv\,,d\mu$. The case of systems is much less investigated. Previous results have been obtained only in the framework of linear symmetric elliptic operators in [18]. Further reference on this subject can be found in the book [9] and in the papers [11] and [16], which contain a wide bibliography.

Our result provides a description of the limiting behavior of sequences of Dirichlet boundary value problems not only for monotone operators of Leray–Lions type, but also covering the case of systems related to linear possibly nonsymmetric operators or nonlinear homogeneous operators, which were not included in previous results. The proof follows the lines of [3], where Theorem 1.1 is obtained in the scalar case, under some additional assumptions on the sequence (Ω_n) which imply in particular that the measure μ in the limit problem is bounded. The idea of the proof is essentially to compare our sequence of problems with a sequence of model problems for which the behavior is known (for instance scalar problems with the p-Laplace operator).

In section 2 we recall some preliminary results and notation and in section 3 we state some known results in the study of the asymptotic behavior of scalar problems with the p-Laplace operator.

Section 4 is dedicated to a careful study of the behavior of a sequence of "correctors" for the p-Laplace operator, as introduced in [16]. In section 5 we state the problem and we prove, following the line of [1] and [16], that a sequence of solutions of problems (1.4) which converges weakly in $W^{1,p}(\Omega, \mathbf{R}^M)$ converges also strongly in $W^{1,r}(\Omega, \mathbf{R}^M)$ for every $r < p$ (see Proposition 5.4). In section 6 we prove the compactness result. In section 7 we prove a correctors result, in the general context of nonlinear monotone vector operators. Indeed, the sequence of gradients (Du_n) of solutions of problems (1.1) converges to Du a priori only weakly in L^p by Theorem 1.1. Hence to obtain a strong convergence it is necessary to add a further sequence which depends only on the limit function u. The construction of such a sequence is provided by Theorem 7.1 and is new also in the case of linear systems. For previous correctors results, see, e.g., [7], [11], [3]. Section 8 is devoted to the analysis of some special cases. In particular we obtain a simpler form for the extra term and for the correctors in the linear case and in the homogeneous case, in agreement with the previous scalar results. The structure of the extra term is proved to depend only on the asymptotic behavior of the function $a(x, \xi)$ for $\xi \to \infty$ (see section 9). In the last section our result is applied also to the treatment of asymptotic problems in a class of pseudomonotone operators. The extension to the general pseudomonotone operators for the scalar case can be found in [5]. Throughout the paper we treat in detail only the case $p \geq 2$. The case $1 \leq p < 2$ can be treated in a similar way, after proper modification on the growth and coerciveness hypotheses for the operator A. The changes in the proofs can easily be performed using Proposition 3.2 of [17].

2. Notation and preliminaries.

Let N and M be two positive integers, $N \geq 2$; by $\mathbf{M}^{M\times N}$ we denote the space of $M \times N$ real matrices.

Let Ω be a bounded open subset of $\mathbf{R}^N$. We denote by $W_0^{1,p}(\Omega, \mathbf{R}^M)$ and $W^{1,p}(\Omega, \mathbf{R}^M)$, $1 < p < +\infty$, the usual Sobolev spaces (of $\mathbf{R}^M$-valued functions) and by $W^{-1,p'}(\Omega, \mathbf{R}^M)$, $1/p' + 1/p = 1$, the dual of $W_0^{1,p}(\Omega, \mathbf{R}^M)$. By $W_c^{1,p}(\Omega, \mathbf{R}^M)$ and $W_{\mathrm{loc}}^{1,p}(\Omega, \mathbf{R}^M)$ we denote respectively the space of all functions in $W^{1,p}(\Omega, \mathbf{R}^M)$ with compact support in Ω and the space of all functions which belong to $W^{1,p}(U, \mathbf{R}^M)$ for every open set $U \subset\subset \Omega$. When $p = 2$ we adopt the standard notation $H^1(\Omega, \mathbf{R}^M)$, $H_0^1(\Omega, \mathbf{R}^M)$, and $H^{-1}(\Omega, \mathbf{R}^M)$.

By $L^p_\mu(\Omega, \mathbf{R}^M)$, $1 \leq p \leq +\infty$, we denote the usual Lebesgue space with respect to the measure μ. If μ is the Lebesgue measure, we use the standard notation $L^p(\Omega, \mathbf{R}^M)$.

When we consider space of scalar functions ($M = 1$), we omit $\mathbf{R}^M$ in the notations above.

Let $u \in W^{1,p}(\Omega)$ and $k \in \mathbf{R}$. By $T_k u$ we shall denote the truncation at the level k which is the function in $W^{1,p}(\Omega)$ defined by $T_k u = (-k) \wedge u \vee k$.

Let A be an open set in $\mathbf{R}^N$, $u : A \to \mathbf{R}^M$ and $a, b \in \mathbf{R}$; we shall denote by $\{a \leq |u| \leq b\}_A$ the set of all $x \in A$ such that $a \leq |u(x)| \leq b$. When $A = \Omega$ we shall omit Ω in the notation above.

We shall use $o_{m,n}$ (respectively, o_n) to denote a sequence of real numbers such that $\lim_{m\to\infty} \lim_{n\to\infty} o_{m,n} = 0$ (respectively, $\lim_{n\to\infty} o_n = 0$).

If $E \subseteq \Omega$, the (*harmonic*) *p-capacity* of E in Ω, denoted by $C_p(E)$, is defined as the infimum of

$$\int_\Omega |Du|^p \, dx$$

over the set of all functions $u \in W_0^{1,p}(\Omega)$ such that $u \geq 1$ almost everywhere (a.e.) in a neighborhood of E.

We say that a property $\mathcal{P}(x)$ holds *p-quasi everywhere* (abbreviated as p-q.e.) in a set E if it holds for all $x \in E$ except for a subset N of E of p-capacity zero. The expression *μ-almost everywhere* (abbreviated as μ-a.e.) refers, as usual, to the analogous property for a Borel measure μ.

A function $u \colon \Omega \to \mathbf{R}^M$ is said to be *p-quasi continuous* if for every $\varepsilon > 0$ there exists a set $A \subseteq \Omega$, with $C_p(A) < \varepsilon$, such that the restriction of u to $\Omega \backslash A$ is continuous.

It is well known that every $u \in W^{1,p}(\Omega, \mathbf{R}^M)$ has a p-quasi continuous representative, which is uniquely defined up to a set of p-capacity zero. In the following we shall always identify u with its p-quasi continuous representative, so that the pointwise values of a function $u \in W^{1,p}(\Omega, \mathbf{R}^M)$ are defined p-q.e. in Ω.

A subset A of Ω is said to be *p-quasi open* in Ω if for every $\varepsilon > 0$ there exists an open subset A_ε of Ω, with $C_p(A_\varepsilon) < \varepsilon$, such that $A \cup A_\varepsilon$ is open. It is easy to see that if a function $u \colon \Omega \to \mathbf{R}$ is p-quasi continuous, then the set $\{u > c\}$ is p-quasi open for every $c \in \mathbf{R}$. For all these properties of p-quasi continuous representatives of Sobolev functions we refer to [28, Chapter 3].

By a nonnegative *Borel measure* in Ω we mean a countably additive set function defined in the Borel σ-field of Ω and with values in $[0, +\infty]$. By a nonnegative *Radon measure* in Ω we mean a nonnegative Borel measure which is finite on every compact subset of Ω. We shall always identify a nonnegative Borel measure with its completion.

We say that a Radon measure ν on Ω belongs to $W^{-1,p'}(\Omega)$ if there exists $f \in W^{-1,p'}(\Omega)$ such that

$$\langle f, \varphi \rangle = \int_\Omega \varphi \, d\nu \qquad \forall \, \varphi \in C_0^\infty(\Omega) \,, \tag{2.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $W^{-1,p'}(\Omega)$ and $W_0^{1,p}(\Omega)$. We shall always identify f and ν. Note that, by the Riesz theorem, for every positive functional $f \in W^{-1,p'}(\Omega)$, there exists a Radon measure ν such that (2.1) holds.

We denote by $\mathcal{M}_0^p(\Omega)$ the class of all Borel measures which vanish on the sets of p-capacity zero and satisfy the following condition:

$$\mu(B) = \inf\{\mu(A) \; : \; A \; p\text{-quasi open}, \; B \subseteq A \subseteq \Omega\}$$

for every Borel set $B \subseteq \Omega$. It is well known that every Radon measure which belongs to $W^{-1,p'}(\Omega)$ belongs also to $\mathcal{M}_0^p(\Omega)$ (see [28, section 4.7]).

3. Preliminary results on the relaxed Dirichlet problem with the p-Laplace operator. Let Ω be a bounded open subset of $\mathbf{R}^N$, $N \geq 2$. Let $2 \leq p < +\infty$ and let $\mu \in \mathcal{M}_0^p(\Omega)$. In the following we shall consider the space $W_0^{1,p}(\Omega) \cap L_\mu^p(\Omega)$ of all functions $u \in W_0^{1,p}(\Omega)$ such that $\int_\Omega |u|^p d\mu < +\infty$. With the norm

$$\|u\|_{W_0^{1,p}(\Omega) \cap L_\mu^p(\Omega)} \; = \left(\int_\Omega |Du|^p dx + \int_\Omega |u|^p d\mu \right)^{\frac{1}{p}}$$

the space $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$ is a reflexive Banach space.

Let f be a functional belonging to $(W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega))'$ (the dual space of $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$) and let us consider the following variational problem:

$$\tag{3.1} \begin{cases} u \in W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega), \\ \displaystyle\int_\Omega |Du|^{p-2} Du Dv \, dx + \int_\Omega |u|^{p-2} uv \, d\mu \; = \; \langle f, v \rangle \\ \forall \, v \in W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega). \end{cases}$$

Since the operator from $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$ to $(W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega))'$ mapping $u \in W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$ to the functional defined by $\int_\Omega |Du|^{p-2} Du Dv \, dx + \int_\Omega |u|^{p-2} uv \, d\mu$ for every $v \in W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$ is a maximal monotone operator and the space $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$ is reflexive, we get that there exists a unique solution u of problem (3.1).

Remark 3.1. It is easy to see that the dual of $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$, $(W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega))'$, coincides with $W^{-1,p'}(\Omega) + L^{p'}_\mu(\Omega)$, so that, in particular, an element of the space $W^{-1,p'}(\Omega)$ can be seen as an element of $(W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega))'$. In what follows, with a slight abuse of notation, $\langle f, v \rangle$ will denote the duality pairing between $(W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega))'$ and $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$, in the general case $f \in (W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega))'$, and the duality pairing between $W^{-1,p'}(\Omega)$ and $W_0^{1,p}(\Omega)$, in the case $f \in W^{-1,p'}(\Omega)$.

Many results similar to those given in the linear case (comparison principle, compactness, etc.) have been proved by Dal Maso and Murat (see [16] and [15]) for nonlinear problems of the type (3.1) (in general for nonlinear homogeneous operators).

PROPOSITION 3.2. *Let f_1, $f_2 \in W^{-1,p'}(\Omega)$ and let μ_1, $\mu_2 \in \mathcal{M}_0^p(\Omega)$. Let u_1, $u_2 \in W_0^{1,p}(\Omega)$ be the solutions of problem* (3.1) *corresponding to f_1, μ_1 and to f_2, μ_2. If $0 \le f_1 \le f_2$ and $\mu_2 \le \mu_1$ in Ω, then $0 \le u_1 \le u_2$ p-q.e. in Ω.*

Proof. See [15, Proposition 2.7]. □

In the space $\mathcal{M}_0^p(\Omega)$ it is possible to introduce a notion of convergence relative to the p-Laplace operator, $\Delta_p u = \operatorname{div}(|Du|^{p-2} Du)$.

DEFINITION 3.3. *Let (μ_n) be a sequence of measures of $\mathcal{M}_0^p(\Omega)$ and let $\mu \in \mathcal{M}_0^p(\Omega)$. We say that (μ_n) $\gamma^{-\Delta_p}$-converges to the measure μ if, for every $f \in W^{-1,p'}(\Omega)$, the sequence (u_n) of solutions of problems*

$$\tag{3.2} \begin{cases} u_n \in W_0^{1,p}(\Omega) \cap L^p_{\mu_n}(\Omega), \\ \displaystyle\int_\Omega |Du_n|^{p-2} Du_n Dv \, dx + \int_\Omega |u_n|^{p-2} u_n v \, d\mu_n \; = \; \langle f, v \rangle \\ \forall \, v \in W_0^{1,p}(\Omega) \cap L^p_{\mu_n}(\Omega) \end{cases}$$

converges weakly in $W_0^{1,p}(\Omega)$ to the solution u of problem (3.1).

THEOREM 3.4. *Every sequence of measures in $\mathcal{M}_0^p(\Omega)$ contains a $\gamma^{-\Delta_p}$-convergent subsequence.*

Proof. See [10, Theorem 2.1] or [15, Theorem 6.5]. □

Many properties of the measure $\mu \in \mathcal{M}_0^p(\Omega)$ can be studied by means of the solution w of the problem

$$\begin{cases} w \in W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega), \\ \displaystyle\int_\Omega |Dw|^{p-2} Dw Dv \, dx + \int_\Omega |w|^{p-2} wv \, d\mu \; = \; \int_\Omega v \, dx \\ \forall \, v \in W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega). \end{cases} \tag{3.3}$$

By the comparison principle (Proposition 3.2), the function w is bounded in $L^\infty(\Omega)$ by a constant which does not depend on μ (see [15, section 2]) and satisfies $w \geq 0$ p-q.e. in Ω.

THEOREM 3.5. *Let $\mu \in \mathcal{M}_0^p(\Omega)$, let w be the solution of problem* (3.3) *and let $\nu = 1 + \Delta_p w$ in the sense of $W^{-1,p'}(\Omega)$. Then ν is a nonnegative Radon measure of $W^{-1,p'}(\Omega)$ and*

$$\nu(B \cap \{w > 0\}) \; = \; \int_B w^{p-1} d\mu \tag{3.4}$$

for every Borel set $B \subseteq \Omega$. Reciprocally, we have

$$\mu(B) \; = \; \begin{cases} \displaystyle\int_B \frac{1}{w^{p-1}} d\nu & \textit{if } C_p(B \cap \{w = 0\}) = 0, \\ +\infty & \textit{if } C_p(B \cap \{w = 0\}) > 0 \end{cases}$$

for any Borel set $B \subseteq \Omega$.

Proof. See [15, Theorem 5.1] and [11, Proposition 3.4] for the linear case. □

The next proposition gives two density results which will be useful in what follows.

PROPOSITION 3.6. *Let $\mu \in \mathcal{M}_0^p(\Omega)$ and let w the solution of problem* (3.3). *Then*

(a) *the set $\{w\psi : \psi \in C_0^\infty(\Omega)\}$ is dense in $W_0^{1,p}(\Omega) \cap L^p_\mu(\Omega)$ and hence in $W^{1,p}_{\text{loc}}(\Omega) \cap L^p_\mu(\Omega)$.*

(b) *the set Λ of all functions of the form $w \sum_{i=1}^l a_i 1_{K_i}$ where $a_i \in \mathbf{R}$ and K_i are closed subsets of Ω such that $w = 0$ μ-a.e. on $K_i \cap K_j$, with $i \neq j$, is dense in $L^p_\mu(\Omega)$.*

Proof. The proof of part (a) is given in [15, Proposition 5.5]. In order to prove part (b), we consider the measure $\lambda = w^p \mu$. Since w belongs to $L^p_\mu(\Omega)$, the measure λ is a Borel bounded measure and therefore the set of all step functions of the form $\sum_{i=1}^l a_i 1_{K_i}$, where $a_i \in \mathbf{R}$ and K_i are closed subsets of Ω such that, for $i \neq j$, $\lambda(K_i \cap K_j) = 0$, is dense in $L^p_\lambda(\Omega)$. If u belongs to $L^p_\mu(\Omega)$, then u/w belongs to $L^p_\lambda(\Omega)$, and for every $\zeta \in \Lambda$ we have

$$\int_\Omega \left| \zeta - \frac{u}{w} \right|^p d\lambda = \int_\Omega |w\zeta - u|^p \, d\mu$$

which gives part (b). □

Finally the solutions of problems (3.3) are useful to characterize the $\gamma^{-\Delta_p}$-convergence in $\mathcal{M}_0^p(\Omega)$. Let (μ_n) be a sequence of measures in $\mathcal{M}_0^p(\Omega)$, and let w_n be the solutions of the problems

$$\begin{cases} w_n \in W_0^{1,p}(\Omega) \cap L^p_{\mu_n}(\Omega), \\ \displaystyle\int_\Omega |Dw_n|^{p-2} Dw_n Dv \, dx + \int_\Omega |w_n|^{p-2} w_n v \, d\mu_n \; = \; \int_\Omega v \, dx \\ \forall \, v \in W_0^{1,p}(\Omega) \cap L^p_{\mu_n}(\Omega). \end{cases} \tag{3.5}$$

The following result characterizes the $\gamma^{-\Delta_p}$-convergence in terms of the convergence of the functions w_n.

Theorem 3.7. *The following conditions are equivalent:*

(a) *(w_n) converges to w weakly in $W_0^{1,p}(\Omega)$;*

(b) *(μ_n) $\gamma^{-\Delta_p}$-converges to μ.*

Proof. See [15, Theorem 6.3] and [11, Theorem 4.3] for the linear case. □

Remark 3.8. If (μ_n) $\gamma^{-\Delta_p}$-converges to μ, then the sequence (w_n) converges to w strongly in $W_0^{1,r}(\Omega)$ for every $1 \le r < p$ and hence, a subsequence of (Dw_n) converges to Dw pointwise a.e. in Ω (see [15, Theorem 6.8]).

4. Sequences in the spaces $W^{1,p}\cap L^p_{\mu_n}$.

In this section (μ_n) will be a sequence of $\mathcal{M}_0^p(\Omega)$ which $\gamma^{-\Delta_p}$-converges to a measure $\mu \in \mathcal{M}_0^p(\Omega)$. We shall use the sequence (w_n) of the solutions of problems (3.5) to investigate the behavior of an arbitrary sequence (u_n), with $u_n \in W^{1,p}(\Omega) \cap L^p_{\mu_n}(\Omega)$, which converges weakly in $W^{1,p}(\Omega)$. By Remark 3.8 we may assume that (w_n) and (Dw_n), respectively, converge to w and Dw pointwise a.e. in Ω.

Let us prove some technical lemmas that will be useful in the remainder of this paper.

Lemma 4.1. *Let Ω' be an open subset of Ω. For every $\varphi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$ we have*

$$\lim_{n\to\infty}\Big(\int_{\Omega'} |Dw_n|^p\varphi\,dx + \int_{\Omega'} |w_n|^p\varphi\,d\mu_n\Big) = \int_{\Omega'} |Dw|^p\varphi\,dx + \int_{\Omega'} |w|^p\varphi\,d\mu. \tag{4.1}$$

Proof. Let $\varphi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$ and let us extend φ to Ω by setting $\varphi = 0$ in $\Omega \setminus \Omega'$. Thus $w_n\varphi$ belongs to $W_0^{1,p}(\Omega)$, and we can take it as test function in (3.3). Therefore by Remark 3.8 we obtain

$$\begin{aligned}
&\lim_{n\to\infty}\Big(\int_{\Omega'} |Dw_n|^p\varphi\,dx + \int_{\Omega'} |w_n|^p\varphi\,d\mu_n\Big)\\
&\quad = \lim_{n\to\infty}\Big(\int_{\Omega'} w_n\varphi\,dx - \int_{\Omega'} |Dw_n|^{p-2}Dw_nD\varphi\,w_n dx\Big)\\
&\quad = \int_{\Omega'} w\varphi\,dx - \int_{\Omega} |Dw|^{p-2}DwD\varphi\,w\,dx = \int_{\Omega'} |Dw|^p\varphi\,dx + \int_{\Omega'} |w|^p\varphi\,d\mu,
\end{aligned}$$

which concludes the proof. □

Lemma 4.2. *Let Ω' be an open subset of Ω. For every $\varphi, \psi \in W^{1,p}(\Omega') \cap L^\infty(\Omega')$, with φ or ψ in $W_0^{1,p}(\Omega')$, we have*

$$\lim_{n\to\infty}\Big(\int_{\Omega'} |D(w_n\psi)|^p\varphi\,dx + \int_{\Omega'} |w_n\psi|^p\varphi\,d\mu_n\Big) = \int_{\Omega'} |D(w\psi)|^p\varphi\,dx + \int_{\Omega'} |w\psi|^p\varphi\,d\mu. \tag{4.2}$$

Proof. Let $\varphi, \psi \in W^{1,p}(\Omega') \cap L^\infty(\Omega')$, with φ or ψ in $W_0^{1,p}(\Omega')$. Since for every $\xi_1, \xi_2 \in \mathbf{R}^N$ and for every $p \ge 2$ the following inequality holds

$$\big||\xi_1|^p - |\xi_2|^p\big| \le p\big(|\xi_1| + |\xi_2|\big)^{p-1}|\xi_1 - \xi_2|, \tag{4.3}$$

we have

$$\big||\psi Dw_n + w_nD\psi|^p - |\psi Dw_n|^p\big| \le p\big(|\psi Dw_n + w_nD\psi| + |\psi Dw_n|\big)^{p-1}|w_nD\psi|,$$

where by Remark 3.8 the left-hand side converges pointwise to $\big||\psi Dw + wD\psi|^p - |\psi Dw|^p\big|$ and the right-hand side is uniformly integrable. Then $(|D(w_n\psi)|^p - |\psi Dw_n|^p)$ converges to $|D(w\psi)|^p - |\psi Dw|^p$ strongly in $L^1(\Omega')$. This implies that

$$\int_{\Omega'} |D(w_n\psi)|^p\varphi\,dx = \int_{\Omega'} |D(w\psi)|^p\varphi\,dx - \int_{\Omega'} |Dw|^p|\psi|^p\varphi\,dx + \int_{\Omega'} |Dw_n|^p|\psi|^p\varphi\,dx + o_n$$

and therefore the conclusion follows from Lemma 4.1. □

Lemma 4.3. *Let Ω' be an open subset of Ω, let $u \in W^{1,p}(\Omega') \cap L^p_\mu(\Omega')$, and let (ψ_m) be a sequence of functions in $C^\infty_0(\Omega')$ such that $(w\psi_m)$ converges to u strongly in $W^{1,p}_{\text{loc}}(\Omega') \cap L^p_\mu(\Omega')$. Then*

$$\lim_{m\to\infty}\lim_{n\to\infty}\Big(\int_{\Omega'} |D(w_n\psi_m - u)|^p\varphi\,dx + \int_{\Omega'} |w_n\psi_m|^p\varphi\,d\mu_n\Big) = \int_{\Omega'} |u|^p\varphi\,d\mu \tag{4.4}$$

for every $\varphi \in W^{1,p}_c(\Omega') \cap L^\infty(\Omega')$.

Proof. As in the proof of Lemma 4.2, $\big(|D(w_n\psi_m - u)|^p - |D(w_n\psi_m)|^p\big)$ converges to $|D(w\psi_m - u)|^p - |D(w\psi_m)|^p$ strongly in $L^1(\Omega')$ as $n \to \infty$. Let $\varphi \in W^{1,p}_c(\Omega') \cap L^\infty(\Omega')$; thus, by Lemma 4.2, we get

$$\begin{aligned}\int_{\Omega'} |D(w_n\psi_m - u)|^p\varphi\,dx + \int_{\Omega'} |w_n\psi_m|^p\varphi\,d\mu_n &= \int_{\Omega'} \big(|D(w_n\psi_m - u)|^p - |D(w_n\psi_m)|^p\big)\varphi\,dx\\ + \int_{\Omega'} |D(w_n\psi_m)|^p\varphi\,dx + \int_{\Omega'} |w_n\psi_m|^p\varphi\,d\mu_n &= \int_{\Omega'} \big(|D(w\psi_m - u)|^p - |D(w\psi_m)|^p\big)\varphi\,dx\\ + \int_{\Omega'} |D(w\psi_m)|^p\varphi\,dx + \int_{\Omega'} |w\psi_m|^p\varphi\,d\mu + o_n &= \int_{\Omega'} |u|^p\varphi\,d\mu + o_{m,n}.\end{aligned}$$

The conclusion follows by taking the limit first as $n \to \infty$ and then as $m \to \infty$. □

Let Ω' be an open subset of Ω. The following theorem shows that if a sequence (u_n), with $u_n \in W^{1,p}(\Omega') \cap L^p_{\mu_n}(\Omega')$, converges weakly in $W^{1,p}(\Omega')$ to a function $u \in W^{1,p}(\Omega')$, and there exists a constant $C > 0$ such that

$$\int_{\Omega'} |u_n|^p d\mu_n \le C \tag{4.5}$$

for every $n \in \mathbf{N}$, then the function u belongs to $L^p_\mu(\Omega')$.

Theorem 4.4. *Let (u_n) be a sequence such that $u_n \in W^{1,p}(\Omega') \cap L^p_{\mu_n}(\Omega')$. Suppose that (u_n) converges weakly in $W^{1,p}(\Omega')$ to some function u. Then*

$$\liminf_{n\to\infty}\Big(\int_{\Omega'} |Du_n|^p dx + \int_{\Omega'} |u_n|^p d\mu_n\Big) \ge \int_{\Omega'} |Du|^p dx + \int_{\Omega'} |u|^p d\mu. \tag{4.6}$$

In particular, if (4.5) *holds, then $u \in W^{1,p}(\Omega') \cap L^p_\mu(\Omega')$.*

The result of Theorem 4.4 can be obtained as a direct consequence of the Γ-convergence of the functionals $\int_{\Omega'} |Du_n|^p dx + \int_{\Omega'} |u_n|^p d\mu_n$ to the functional $\int_{\Omega'} |Du|^p dx + \int_{\Omega'} |u|^p d\mu$ proved in [10]. For the sake of completeness we shall give an alternative proof of Theorem 4.4 which does not involve Γ-convergence theory. Before proving Theorem 4.4, let us prove two preliminary lemmas.

Lemma 4.5. *Let (u_n) be a sequence such that $u_n \in W^{1,p}(\Omega') \cap L^p_{\mu_n}(\Omega')$ and such that* (4.5) *holds. Suppose that (u_n) converges weakly in $W^{1,p}(\Omega')$ to some function u. Then $\{u = 0\}_{\Omega'} \supseteq \{w = 0\}_{\Omega'}$.*

Proof. Taking into account the decomposition $u_n = u_n^+ - u_n^-$, where u_n^+ and u_n^- denote respectively the positive and the negative part of u_n, it is not restrictive to assume $u_n \geq 0$ p-q.e. in Ω'.

We shall prove first the result in the special case where the functions u_n and u belong to $W_0^{1,p}(\Omega')$, and we shall suppose, also, that there exists a constant $K > 0$ such that $u_n \leq K$ p-q.e. in Ω' and hence $u \leq K$ p-q.e. in Ω'.

For every $m \in \mathbf{N}$ let us consider the sequence (u_n^m) of the solutions of the problems

$$\text{(4.7)} \qquad \begin{cases} u_n^m \in W_0^{1,p}(\Omega') \cap L^p_{\mu_n}(\Omega'), \\ \displaystyle\int_{\Omega'} |Du_n^m|^{p-2} Du_n^m Dv\,dx + \int_{\Omega'} |u_n^m|^{p-2} u_n^m v\,d\mu_n \\ \qquad = m \displaystyle\int_{\Omega'} \big(|u_n|^{p-2}u_n - |u_n^m|^{p-2}u_n^m\big) v\,dx \\ \forall\, v \in W_0^{1,p}(\Omega') \cap L^p_{\mu_n}(\Omega'), \end{cases}$$

which, extended to Ω by setting $u_n^m = 0$ in $\Omega \backslash \Omega'$, are also the solutions of the following equivalent problems:

$$\text{(4.8)} \qquad \begin{cases} u_n^m \in W_0^{1,p}(\Omega) \cap L^p_{\hat{\mu}_n}(\Omega), \\ \displaystyle\int_{\Omega} |Du_n^m|^{p-2} Du_n^m Dv\,dx + \int_{\Omega} |u_n^m|^{p-2} u_n^m v\,d\hat{\mu}_n \\ \qquad = m \displaystyle\int_{\Omega} \big(|u_n|^{p-2}u_n - |u_n^m|^{p-2}u_n^m\big) v\,dx \\ \forall\, v \in W_0^{1,p}(\Omega) \cap L^p_{\hat{\mu}_n}(\Omega), \end{cases}$$

where $\hat{\mu}_n$ is the measure defined by

$$\hat{\mu}_n(B) = \begin{cases} \mu_n(B) & \text{if } C_p(B \cap (\Omega \setminus \Omega')) = 0, \\ +\infty & \text{if } C_p(B \cap (\Omega \setminus \Omega')) > 0 \end{cases}$$

for any Borel set $B \subseteq \Omega$. By the comparison principle (Proposition 3.2) we have that

$$\text{(4.9)} \qquad 0 \leq u_n^m \leq m^{\frac{1}{p-1}} K w_n \qquad p\text{-q.e. in } \Omega.$$

By taking in (4.7) $u_n^m - u_n$ as a test function we get

$$\text{(4.10)} \qquad \begin{aligned} &\int_{\Omega'} \big(|Du_n^m|^{p-2}Du_n^m - |Du_n|^{p-2}Du_n\big) D(u_n^m - u_n)\,dx \\ &\quad + \int_{\Omega'} \big(|u_n^m|^{p-2}u_n^m - |u_n|^{p-2}u_n\big)(u_n^m - u_n)\,d\mu_n \\ &\quad + m \int_{\Omega'} \big(|u_n^m|^{p-2}u_n^m - |u_n|^{p-2}u_n\big)(u_n^m - u_n)\,dx \\ = &- \int_{\Omega'} |Du_n|^{p-2}Du_n D(u_n^m - u_n)\,dx - \int_{\Omega'} |u_n|^{p-2}u_n(u_n^m - u_n)\,d\mu_n. \end{aligned}$$

Since for every $\xi_1, \xi_2 \in \mathbf{R}^N$ and for every $p \geq 2$ we have

$$\text{(4.11)} \qquad (|\xi_1|^{p-2}\xi_1 - |\xi_2|^{p-2}\xi_2)(\xi_1 - \xi_2) \geq 2^{2-p}|\xi_1 - \xi_2|^p,$$

applying Young's inequality in (4.10) we get

$$\begin{aligned}2^{2-p}\int_{\Omega'}|D(u_n^m-u_n)|^p dx+2^{2-p}\int_{\Omega'}|u_n^m-u_n|^p d\mu_n+2^{2-p}m\int_{\Omega'}|u_n^m-u_n|^p dx\\ \le\frac{1}{\varepsilon^{p'}p'}\Big(\int_{\Omega'}|Du_n|^p dx+\int_{\Omega'}|u_n|^p d\mu_n\Big)\\ +\frac{\varepsilon^p}{p}\Big(\int_{\Omega'}|D(u_n^m-u_n)|^p dx+\int_{\Omega'}|u_n^m-u_n|^p d\mu_n\Big),\end{aligned}$$

where $\varepsilon>0$ is an arbitrary real number. Since (u_n) is bounded in $W_0^{1,p}(\Omega')$ and (4.5) holds, by choosing ε small enough we obtain that there exists a constant $C>0$ such that

$$\int_{\Omega'}|D(u_n^m-u_n)|^p dx+m\int_{\Omega'}|u_n^m-u_n|^p dx\ \le\ C. \tag{4.12}$$

By (4.12) we have that the sequence (u_n^m) is bounded in $W_0^{1,p}(\Omega')$, uniformly in m and n. Then for every $m\in\mathbf{N}$ there exists a subsequence of (u_n^m) (we can choose the subsequence independent of m) which converges to a function u^m weakly in $W_0^{1,p}(\Omega')$. By the weak lower semicontinuity of the norm, the sequence (u^m) is also bounded in $W_0^{1,p}(\Omega')$. Moreover by (4.12) we get

$$\int_{\Omega'}|u^m-u|^p dx\ =\ \lim_{n\to\infty}\int_{\Omega'}|u_n^m-u_n|^p dx\ \le\ \frac{C}{m},$$

and hence (u^m) converges weakly to u in $W_0^{1,p}(\Omega')$. By (4.9) we have that $|u^m|\le m^{1/(p-1)}Kw$ p-q.e. in Ω' and hence u^m belongs to the set $\mathcal{K}=\{v\in W_0^{1,p}(\Omega'):v=0\ p\text{-q.e. in }\{w=0\}_{\Omega'}\}$. Since $\mathcal{K}$ is convex and closed in $W_0^{1,p}(\Omega')$, it is weakly closed. Therefore $u\in\mathcal{K}$ and hence $\{u=0\}_{\Omega'}\supseteq\{w=0\}_{\Omega'}$.

Finally let us consider the general case where the sequence (u_n) is not bounded in $L^\infty(\Omega')$ but $u_n\in W^{1,p}(\Omega')\cap L^p_{\mu_n}(\Omega')$, satisfies (4.5), and converges weakly in $W^{1,p}(\Omega')$ to u. Let Φ be a function in $W_0^{1,\infty}(\Omega')$, with $\Phi>0$ in Ω', and for every $n\in\mathbf{N}$ let T_1u_n be the truncation at the level 1 of u_n. Since $\Phi T_1u_n\in W_0^{1,p}(\Omega')\cap L^p_{\mu_n}(\Omega')$ and the sequence (ΦT_1u_n) satisfies (4.5), is bounded in $L^\infty(\Omega')$, and converges weakly in $W_0^{1,p}(\Omega')$ to ΦT_1u, by the previous step we can conclude that $\{\Phi T_1u=0\}_{\Omega'}\supseteq\{w=0\}_{\Omega'}$ and hence $\{u=0\}_{\Omega'}\supseteq\{w=0\}_{\Omega'}$. □

Lemma 4.6. *Let (v_n), with $v_n\in W^{1,p}(\Omega')\cap L^p_{\mu_n}(\Omega')$, be a sequence which converges to a function v weakly in $W^{1,p}(\Omega')$, and suppose that there exists a constant $C>0$ such that*

$$\int_{\Omega'}|v_n|^p d\mu_n\ \le\ C \tag{4.13}$$

for every $n\in\mathbf{N}$. Then we have

$$\begin{aligned}\lim_{n\to\infty}\Big(\int_{\Omega'}\varphi|D(w_n\psi)|^{p-2}D(w_n\psi)Dv_n\,dx+\int_{\Omega'}\varphi|w_n\psi|^{p-2}w_n\psi v_n\,d\mu_n\Big)\\ =\int_{\Omega'}\varphi|D(w\psi)|^{p-2}D(w\psi)Dv\,dx+\int_{\Omega'}\varphi|w\psi|^{p-2}w\psi v\,d\mu\end{aligned} \tag{4.14}$$

for every $\psi\in W^{1,p}(\Omega')\cap L^\infty(\Omega')$ and for every $\varphi\in W_0^{1,p}(\Omega')\cap L^\infty(\Omega')$.

Proof. Let $\psi \in W^{1,p}(\Omega') \cap L^\infty(\Omega')$ and $\varphi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$. Since for every $p \geq 2$ the following inequality holds

$$\big||\xi_1|^{p-2}\xi_1 - |\xi_2|^{p-2}\xi_2\big| \leq (p-1)(|\xi_1| + |\xi_2|)^{p-2}|\xi_1 - \xi_2| \tag{4.15}$$

for every $\xi_1, \xi_2 \in \mathbf{R}^N$, as in Lemma 4.2, we can conclude that $\big(|D(w_n\psi)|^{p-2}D(w_n\psi) - |\psi Dw_n|^{p-2}\psi Dw_n\big)$ converges to $|D(w\psi)|^{p-2}D(w\psi) - |\psi Dw|^{p-2}\psi Dw$ strongly in $L^{p'}(\Omega', \mathbf{R}^N)$. Thus

$$\begin{aligned} &\lim_{n\to\infty}\Big(\int_{\Omega'} \varphi|D(w_n\psi)|^{p-2}D(w_n\psi)Dv_n\,dx - \int_{\Omega'}\varphi|\psi Dw_n|^{p-2}\psi Dw_n Dv_n\,dx\Big) \\ &\qquad = \int_{\Omega'}\big(|D(w\psi)|^{p-2}D(w\psi) - |\psi Dw|^{p-2}\psi Dw\big)Dv\,\varphi\,dx. \end{aligned} \tag{4.16}$$

In order to conclude the proof it is enough to show that

$$\begin{aligned} &\lim_{n\to\infty}\Big(\int_{\Omega'}\varphi|\psi Dw_n|^{p-2}\psi Dw_n Dv_n\,dx + \int_{\Omega'}\varphi|w_n\psi|^{p-2}w_n\psi v_n\,d\mu_n\Big) \\ &\qquad = \int_{\Omega'}\varphi|\psi Dw|^{p-2}\psi Dw Dv\,dx + \int_{\Omega'}\varphi|\psi|^{p-2}\psi v\,d\nu, \end{aligned} \tag{4.17}$$

where $\nu \in W^{-1,p'}(\Omega)$ is the Radon measure defined in Theorem 3.5. Indeed, since by Lemma 4.5 we have that $\{v = 0\}_{\Omega'} \supseteq \{w = 0\}_{\Omega'}$, by (3.4) we get

$$\int_{\Omega'}\varphi|\psi|^{p-2}\psi v\,d\nu = \int_{\{w>0\}_{\Omega'}}\varphi|\psi|^{p-2}\psi v\,d\nu = \int_{\Omega'}\varphi w^{p-1}|\psi|^{p-2}\psi v\,d\mu;$$

therefore the conclusion follows from (4.16) and (4.17).

It remains to prove (4.17). Let us consider $\phi \in W_0^{1,\infty}(\Omega')$. Taking $\phi v_n \in W_0^{1,p}(\Omega') \cap L^p_{\mu_n}(\Omega') \subset W_0^{1,p}(\Omega) \cap L^p_{\mu_n}(\Omega)$ as a test function in problem (3.5), and taking into account that $\nu = 1 + \Delta_p w$ in $W^{-1,p'}(\Omega)$ (Theorem 3.5), we obtain

$$\begin{aligned} &\lim_{n\to\infty}\int_{\Omega'}\phi|Dw_n|^{p-2}Dw_n Dv_n\,dx + \int_{\Omega'}\phi|w_n|^{p-2}w_n v_n\,d\mu_n \\ &\qquad = \lim_{n\to\infty}\int_{\Omega'}\phi v_n\,dx - \int_{\Omega'}|Dw_n|^{p-2}Dw_n D\phi\, v_n\,dx \\ &= \int_{\Omega'}\phi v\,dx - \int_{\Omega'}|Dw|^{p-2}Dw D\phi\, v\,dx = \int_{\Omega'}\phi|Dw|^{p-2}Dw Dv\,dx + \int_{\Omega'}\phi v\,d\nu. \end{aligned} \tag{4.18}$$

We have to prove that (4.18) holds for every $\phi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$. Let $\phi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$. Since ν is a Radon measure in $W^{-1,p'}(\Omega)$, it is possible to construct a sequence (ϕ_m) of functions in $W_0^{1,\infty}(\Omega')$ bounded in $L^\infty(\Omega')$, which converges to ϕ a.e. and ν-a.e. in Ω'. By (4.18) we have

$$\begin{aligned} &\Big|\int_{\Omega'}\phi|Dw_n|^{p-2}Dw_n Dv_n\,dx + \int_{\Omega'}\phi|w_n|^{p-2}w_n v_n\,d\mu_n - \int_{\Omega'}\phi|Dw|^{p-2}Dw Dv\,dx - \int_{\Omega'}\phi v\,d\nu\Big| \\ &\qquad \leq \Big|\int_{\Omega'}(\phi - \phi_m)|Dw_n|^{p-2}Dw_n Dv_n\,dx + \int_{\Omega'}(\phi - \phi_m)|w_n|^{p-2}w_n v_n\,d\mu_n\Big| \\ &\qquad\qquad + \Big|\int_{\Omega'}(\phi_m - \phi)|Dw|^{p-2}Dw Dv\,dx + \int_{\Omega'}(\phi_m - \phi)v\,d\nu\Big| + o_n. \end{aligned} \tag{4.19}$$

By the dominated convergence theorem the limit as $m \to \infty$ of the second term in the right-hand side of (4.19) is zero. It remains to estimate the first term of the right-hand side of (4.19). Since (ϕ_m) is bounded in $L^\infty(\Omega')$, by Hölder's inequality, (4.13), and Lemma 4.1 we obtain

$$\begin{aligned} &\Big|\int_{\Omega'} (\phi - \phi_m)|Dw_n|^{p-2} Dw_n Dv_n\, dx + \int_{\Omega'} (\phi - \phi_m)|w_n|^{p-2} w_n v_n\, d\mu_n\Big| \\ &\le C\Big(\int_{\Omega'} |Dw_n|^p |\phi - \phi_m|^{\frac{p}{(p-1)}}\, dx + \int_{\Omega'} |w_n|^p |\phi - \phi_m|^{\frac{p}{(p-1)}}\, d\mu_n\Big)^{\frac{p-1}{p}} \\ &= C\Big(\int_{\Omega'} |Dw|^p |\phi - \phi_m|^{\frac{p}{(p-1)}}\, dx + \int_{\Omega'} |w|^p |\phi - \phi_m|^{\frac{p}{(p-1)}}\, d\mu\Big)^{\frac{p-1}{p}} + o_n = o_{m,n}, \end{aligned}$$

where C is a positive constant independent of n and m and where for the last limit we used the dominated convergence theorem. Finally (4.17) follows immediately from (4.18) by choosing $\phi = \varphi|\psi|^{p-2}\psi$. □

We are now in a position to prove Theorem 4.4.

Proof of Theorem 4.4. If $\liminf_{n\to\infty} \int_{\Omega'} |u_n|^p d\mu_n = +\infty$, then inequality (4.6) is trivially satisfied; otherwise it is not restrictive to suppose that (4.5) holds. Let $\psi \in W^{1,p}(\Omega') \cap L^\infty(\Omega')$, and let $\varphi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$ with $\varphi \ge 0$. Since for every $\xi_1, \xi_2 \in \mathbf{R}^N$, by the convexity of the function $|\cdot|^p$, the following inequality holds:

$$|\xi_1|^p - |\xi_2|^p \ \ge\ p|\xi_2|^{p-2}\xi_2(\xi_1 - \xi_2), \tag{4.20}$$

we have

$$\begin{aligned} &\int_{\Omega'} \varphi|Du_n|^p dx + \int_{\Omega'} \varphi|u_n|^p d\mu_n \ \ge\ \int_{\Omega'} \varphi|D(w_n\psi)|^p dx + \int_{\Omega'} \varphi|w_n\psi|^p d\mu_n \\ &+ p\int_{\Omega'} |D(w_n\psi)|^{p-2} D(w_n\psi) D(u_n - w_n\psi)\varphi\, dx + p\int_{\Omega'} |w_n\psi|^{p-2} w_n\psi(u_n - w_n\psi)\varphi\, d\mu_n. \end{aligned}$$

By Lemmas 4.2 and 4.6 we get

$$\begin{aligned} &\liminf_{n\to\infty}\Big(\int_{\Omega'} \varphi|Du_n|^p dx + \int_{\Omega'} \varphi|u_n|^p d\mu_n\Big) \\ &\ge \int_{\Omega'} \varphi|D(w\psi)|^p dx + \int_{\Omega'} \varphi|w\psi|^p d\mu \\ &+ p\int_{\Omega'} |D(w\psi)|^{p-2} D(w\psi) D(u - w\psi)\varphi\, dx + p\int_{\Omega'} |w\psi|^{p-2} w\psi(u - w\psi)\varphi\, d\mu. \end{aligned} \tag{4.21}$$

Assume that $u \in L^\infty(\Omega')$. Let $\varepsilon > 0$ and let us choose in (4.21) $\psi = \frac{u}{w \vee \varepsilon}$ and $\varphi = \phi R_\varepsilon(w)$, with $0 \le \phi \le 1$, $\phi \in W_0^{1,p}(\Omega') \cap L^\infty(\Omega')$, and $R_\varepsilon : \mathbf{R} \mapsto \mathbf{R}$ defined by

$$R_\varepsilon(s) = \begin{cases} 0 & \text{if } s \le \varepsilon, \\ \frac{s}{\varepsilon} - 1 & \text{if } \varepsilon \le s \le 2\varepsilon, \\ 1 & \text{if } 2\varepsilon \le s < +\infty. \end{cases}$$

Since $w\psi = u$ p-q.e. in $\{w > \varepsilon\}$ and $\phi R_\varepsilon(w) = 0$ p-q.e. in $\{w \le \varepsilon\}$, by (4.21) we have

$$\begin{aligned} &\liminf_{n\to\infty}\Big(\int_{\Omega'} |Du_n|^p dx + \int_{\Omega'} |u_n|^p d\mu_n\Big) \\ &\ge \int_{\Omega' \cap \{w > \varepsilon\}} R_\varepsilon(w)\phi|Du|^p dx + \int_{\Omega' \cap \{w > \varepsilon\}} R_\varepsilon(w)\phi|u|^p d\mu, \end{aligned}$$

which, by the monotone convergence theorem, implies

$$\text{(4.22)}\qquad \begin{aligned}&\liminf_{n\to\infty}\Big(\int_{\Omega'}|Du_n|^p dx+\int_{\Omega'}|u_n|^p d\mu_n\Big)\\ &\qquad\geq\int_{\Omega'\cap\{w>0\}}\phi|Du|^p dx+\int_{\Omega'\cap\{w>0\}}\phi|u|^p d\mu\end{aligned}$$

for every $\phi\in W_0^{1,p}(\Omega')\cap L^\infty(\Omega')$ with $0\leq\phi\leq 1$. Since $Du=0$ a.e. in $\{u=0\}_{\Omega'}$ and by Lemma 4.5 $\{u=0\}_{\Omega'}\supseteq\{w=0\}_{\Omega'}$, estimate (4.23) may be written as

$$\liminf_{n\to\infty}\Big(\int_{\Omega'}|Du_n|^p dx+\int_{\Omega'}|u_n|^p d\mu_n\Big)\;\geq\;\int_{\Omega'}\phi|Du|^p dx+\int_{\Omega'}\phi|u|^p d\mu.$$

Thus, by the monotone convergence theorem, we deduce that $u\in L^p_\mu(\Omega')$ and (4.6) holds. If u does not belong to $L^\infty(\Omega')$, it is enough to apply the previous step to the sequence of truncations $T_k(u_n)$ with $k\in\mathbf{N}$. Then we have

$$\begin{aligned}\liminf_{n\to\infty}\Big(\int_{\Omega'}|Du_n|^p dx+\int_{\Omega'}|u_n|^p d\mu_n\Big)\;&\geq\;\liminf_{n\to\infty}\Big(\int_{\Omega'}|DT_k(u_n)|^p dx+\int_{\Omega'}|T_k(u_n)|^p d\mu_n\Big)\\ &\geq\;\int_{\Omega'}|DT_k(u)|^p dx+\int_{\Omega'}|T_k(u)|^p d\mu.\end{aligned}$$

We conclude the proof by the monotone convergence theorem, taking the limit as $k\to\infty$. □

5. Relaxed Dirichlet problems with monotone operators. Let A be the monotone operator defined from $W^{1,p}(\Omega,\mathbf{R}^M)$ to $W^{-1,p'}(\Omega,\mathbf{R}^M)$, with $2\leq p<+\infty$ and $M\geq 2$, mapping $u\in W^{1,p}(\Omega,\mathbf{R}^M)$ in $Au=-\mathrm{div}\big(a(x,Du)\big)\in W^{-1,p'}(\Omega,\mathbf{R}^M)$, where $a\colon\Omega\times\mathbf{M}^{M\times N}\mapsto\mathbf{M}^{M\times N}$ is a Carathéodory function. We shall assume that the function a satisfies the following conditions:

(i) there exists a constant $\alpha>0$ such that

$$(a(x,\xi_1)-a(x,\xi_2))(\xi_1-\xi_2)\geq\alpha|\xi_1-\xi_2|^p$$

for every $\xi_1,\xi_2\in\mathbf{M}^{M\times N}$ and for a.e. $x\in\Omega$;

(ii) there exists a constant $\beta>0$ and a function $h\in L^{\frac{p}{p-2}}(\Omega)$ such that

$$|a(x,\xi_1)-a(x,\xi_2)|\;\leq\;\beta(h(x)+\big(|\xi_1|+|\xi_2|)^{p-2}\big)|\xi_1-\xi_2|$$

for every $\xi_1,\xi_2\in\mathbf{M}^{M\times N}$ and for a.e. $x\in\Omega$;

(iii) $a(x,0)=0$ a.e. in Ω.

These hypotheses imply in particular that the following conditions hold:

(iv) there exists a constant $\eta>0$ and a function $k\in L^{p'}(\Omega)$ such that

$$|a(x,\xi)|\leq k(x)+\eta|\xi|^{p-1}$$

for every $\xi\in\mathbf{M}^{M\times N}$ and for a.e. $x\in\Omega$;

(v) $a(x,\xi)\xi\;\geq\;\alpha|\xi|^p$ for every $\xi\in\mathbf{M}^{M\times N}$ and a.e. $x\in\Omega$.

We shall see in section 10 that the results proved in what follows hold for a class of operators which satisfy more general conditions than (i)–(iv) above.

Given three positive constants c_1, c_2, and σ, with $0<\sigma\leq 1$, let us define the class $\mathcal{F}(c_1,c_2,\sigma)$ of all vector functions $F\colon\Omega\times\mathbf{R}^M\mapsto\mathbf{R}^M$ such that the following properties are satisfied:

(I) $F(x,s)$ is a Borel function;

(II) for every $s_1, s_2 \in \mathbf{R}^M$ and for every $x \in \Omega$ we have

$$(F(x,s_1) - F(x,s_2))(s_1 - s_2) \geq c_1 |s_1 - s_2|^p;$$

(III) for every $s_1, s_2 \in \mathbf{R}^M$ and for every $x \in \Omega$ we have

$$|F(x,s_1) - F(x,s_2)| \leq c_2(|s_1| + |s_2|)^{p-1-\sigma}|s_1 - s_2|^\sigma;$$

(IV) $F(x,0) = 0$ for every $x \in \Omega$.

As consequence of properties (III) and (IV) we have that the function F also satisfies

(V) $|F(x,s)| \leq c_2|s|^{p-1}$ for every $s \in \mathbf{R}^M$ and for every $x \in \Omega$,

and by properties (II) and (IV) we get

(VI) $F(x,s)s \geq c_1|s|^p$ for every $s \in \mathbf{R}^M$ and for every $x \in \Omega$.

In the following we shall fix a constant $L > 0$ and we shall denote by $\mathcal{F}(L)$ the class $\mathcal{F}(\alpha, L, 1)$, where α is the positive constant which appears in condition (i) above.

From now on by C we shall denote a positive constant, depending only on α, β, L, and p, which can change from line to line.

Let $f \in W^{-1,p'}(\Omega, \mathbf{R}^M)$, let (μ_n) be a sequence of $\mathcal{M}_0^p(\Omega)$, and let $F_n \in \mathcal{F}(L)$. Let us consider the following nonlinear systems with boundary Dirichlet condition:

$$\text{(5.1)} \qquad \begin{cases} u_n \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M), \\ \displaystyle\int_\Omega a(x, Du_n)Dv\,dx + \int_\Omega F_n(x,u_n)v\,d\mu_n \;=\; \langle f, v\rangle \\ \forall\, v \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M). \end{cases}$$

Since by Remark 3.1 $\langle f, \cdot\rangle$ is a functional in $(W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M))'$, by assumptions (i)–(v) and (I)–(VI) the theory of monotone operators (see [23]) assures the existence of a unique solution u_n of problem (5.1). From (v) and (VI), taking u_n as a test function in (5.1), it is easy to see that the sequence (u_n) is bounded in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ for any choice of (μ_n) and (F_n). Thus, up to a subsequence, the sequence (u_n) converges weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ to some function $u \in W_0^{1,p}(\Omega, \mathbf{R}^M)$. Our goal is to find the variational problem satisfied by the function u. To this aim we shall consider special sequences of test functions $v_n \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M)$ which converge weakly to some function $v \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M)$, and we shall try to take the limit in problem (5.1). This is the energy method of L. Tartar.

In order to prove that the structure of the limit problem is local (i.e., it does not depend on the choice of the domain Ω and of the boundary data), in what follows, we shall consider a more general situation. Namely, we shall study the asymptotic behavior of an arbitrary sequence (u_n) of solutions of the problems

$$\text{(5.2)} \qquad \begin{cases} u_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du_n)Dv\,dx + \int_{\Omega'} F_n(x,u_n)v\,d\mu_n \;=\; \langle f_n, v\rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \end{cases}$$

where Ω' is an open subset of Ω, $f_n \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))'$. We do not require any boundary data for u_n, while we assume that the sequence (u_n) is bounded in $W^{1,p}(\Omega',\mathbf{R}^M)$, which implies in particular that, up to a subsequence, (u_n) converges weakly to some u in $W^{1,p}(\Omega',\mathbf{R}^M)$. For the sequence (f_n), we shall assume a notion of convergence specified by the following definition.

Definition 5.1. *Let (μ_n) be a sequence of $\mathcal{M}_0^p(\Omega)$ which $\gamma^{-\Delta_p}$-converges to a measure μ. Let (f_n) be a sequence of functionals, with $f_n \in (W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_{\mu_n}(\Omega',\mathbf{R}^M))'$, and let $f \in (W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_\mu(\Omega',\mathbf{R}^M))'$. We shall say that the sequence (f_n) converges to f in the sense of $(\mathcal{H}_{\Omega'})$ if the following condition is satisfied:*

$(\mathcal{H}_{\Omega'})$ *If $v \in W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_\mu(\Omega',\mathbf{R}^M)$, $v_n \in W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_{\mu_n}(\Omega',\mathbf{R}^M)$ for every n, and (v_n) converges to v weakly in $W_0^{1,p}(\Omega',\mathbf{R}^M)$, then $\langle f_n, v_n\rangle \to \langle f, v\rangle$.*

The next lemma gives an estimate of the norm in $(W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_{\mu_n}(\Omega',\mathbf{R}^M))'$ of a sequence of functionals (f_n) which converges in the sense of $(\mathcal{H}_{\Omega'})$, while Proposition 5.3 gives a local estimate of the norm in $L^p_{\mu_n}(\Omega',\mathbf{R}^M)$ of the corresponding solutions u_n of problem (5.2).

Lemma 5.2. *Let $f \in (W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_\mu(\Omega',\mathbf{R}^M))'$, and let $f_n \in (W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_{\mu_n}(\Omega',\mathbf{R}^M))'$ for every n. If (f_n) converges to f in the sense of $(\mathcal{H}_{\Omega'})$, then $(\|f_n\|)$ converges to $\|f\|$, where the norm of f_n (resp., f) is taken in the space $(W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_{\mu_n}(\Omega',\mathbf{R}^M))'$ (resp., $(W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_\mu(\Omega',\mathbf{R}^M))'$).*

Proof. Let (ζ_n) be a sequence such that $\zeta_n \in W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_{\mu_n}(\Omega',\mathbf{R}^M)$, with unit norm and $\|f_n\| = \langle f_n,\zeta_n\rangle$. Then, up to a subsequence, (ζ_n) converges weakly in $W_0^{1,p}(\Omega',\mathbf{R}^M)$ to some function ζ, by Theorem 4.4 $\zeta \in W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_\mu(\Omega',\mathbf{R}^M)$, and $\|\zeta\| \le 1$. Since (f_n) converges in the sense of $(\mathcal{H}_{\Omega'})$ we have that

$$\lim_{n\to\infty} \|f_n\| \,=\, \lim_{n\to\infty}\langle f_n,\zeta_n\rangle \,=\, \langle f,\zeta\rangle \,\le\, \|f\|.$$

In order to prove the opposite inequality let us consider the function ζ such that $\zeta \in W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_\mu(\Omega',\mathbf{R}^M)$, with unit norm, $\|f\| = \langle f,\zeta\rangle$, and let (ψ_m) be a sequence in $C_0^\infty(\Omega',\mathbf{R}^M)$ such that $(w\psi_m)$ converges strongly to ζ in $W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_\mu(\Omega',\mathbf{R}^M)$. By Lemma 4.2 we have that the norm in the space $W_0^{1,p}(\Omega',\mathbf{R}^M) \cap L^p_{\mu_n}(\Omega',\mathbf{R}^M)$ of the functions $w_n\psi_m$ converges to the norm of $w\psi_m$ in the space $W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_\mu(\Omega',\mathbf{R}^M)$. Thus, since $(w_n\psi_m)$ converges weakly in $W_0^{1,p}(\Omega',\mathbf{R}^M)$ to $w\psi_m$, we have

$$\begin{aligned}\|f\| \,=\, \langle f,\zeta\rangle \,&=\, \lim_{m\to\infty}\lim_{n\to\infty}\langle f_n, w_n\psi_m\rangle \\ \le\, \lim_{m\to\infty}\liminf_{n\to\infty}\|f_n\|\,\|w_n\psi_m\| \,&=\, \liminf_{n\to\infty}\|f_n\|\,\|\zeta\| \,=\, \liminf_{n\to\infty}\|f_n\|. \qquad \square\end{aligned}$$

Proposition 5.3. *Let (u_n) be a sequence of solutions of problems* (5.2). *If the sequence (u_n) is bounded in $W^{1,p}(\Omega',\mathbf{R}^M)$, then*

$$\int_{\Omega'} |u_n|^p\varphi\, d\mu_n \,\le\, M \tag{5.3}$$

for every $\varphi \in C_0^1(\Omega')$, with $\varphi \ge 0$, where the constant M depends on the norm in $C_0^1(\Omega')$ of φ.

Proof. The proof follows immediately, taking $u_n\varphi$ as test function in (5.2), by Lemma 5.2 and conditions (v) and (VI). $\square$

The following proposition shows that, without any additional assumption, the sequence (u_n) converges strongly in $W^{1,r}(\Omega',\mathbf{R}^M)$ for every $1 \le r < p$.

Proposition 5.4. *Let (u_n), with $u_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M)$, be a sequence which converges to some function u weakly in $W^{1,p}(\Omega', \mathbf{R}^M)$. Suppose that there exists a sequence (f_n), with $f_n \in (W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))'$, which converges to $f \in (W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$, such that u_n satisfies problem* (5.2). *Then (u_n) converges to u strongly in $W^{1,r}(\Omega', \mathbf{R}^M)$ for every $r < p$, and hence a subsequence of (Du_n) converges to Du pointwise a.e. in Ω'.*

Proof. The proof follows the lines of the one given in [16] (see also [1]).

In the course this proof we shall denote by C a positive constant independent on n. Let $\Psi : \mathbf{R} \mapsto \mathbf{R}$ be a C^1 function which satisfies the following properties:

$$\begin{aligned} &\Psi(t) = 1 \quad \text{if } |t| < 1, \quad \Psi(t) = 0 \quad \text{if } |t| \geq 2, \\ &|\Psi(t)| \leq 1 \quad \forall t \in \mathbf{R}, \quad |\Psi'(t)| \leq 2 \quad \forall t \in \mathbf{R}, \end{aligned}$$

and let $\Phi(y) = \Psi(|y|)y$. Let $\delta > 0$ and, for every $n \in \mathbf{N}$, let $\delta_n \leq \delta$ be a positive real number that we shall fix later. For every such a δ_n we define the function $\Phi_{\delta_n}(y) = \delta_n \Phi(y/\delta_n)$. Given $\varphi \in C^1_0(\Omega')$, with $\varphi \geq 0$, we can take $(\Phi_{\delta_n}(u_n) + w_n \Phi_{\delta_n}(u_n - u))\varphi$ as test function in problem (5.2), and we obtain

$$\begin{aligned} &\int_{\Omega'} a(x, Du_n) D\Phi_{\delta_n}(u_n) Du_n \varphi \, dx \\ &+ \int_{\Omega'} \big(a(x, Du_n) - a(x, Du)\big) D\Phi_{\delta_n}(u_n - u) D(u_n - u) w_n \varphi \, dx \\ &+ \int_{\Omega'} a(x, Du_n) \big(\Phi_{\delta_n}(u_n) \otimes D\varphi + \Phi_{\delta_n}(u_n - u) \otimes D(w_n \varphi)\big) \, dx \\ &+ \int_{\Omega'} F_n(x, u_n)(w_n \Phi_{\delta_n}(u_n - u) + \Phi_{\delta_n}(u_n))\varphi \, d\mu_n \\ &= \langle f_n, (w_n \Phi_{\delta_n}(u_n - u) + \Phi_{\delta_n}(u_n))\varphi \rangle \\ &- \int_{\Omega'} a(x, Du) D\Phi_{\delta_n}(u_n - u) D(u_n - u) w_n \varphi \, dx. \end{aligned} \tag{5.4}$$

Since (u_n) is bounded in $W^{1,p}(\Omega', \mathbf{R}^M)$, (w_n) is bounded in $W^{1,p}(\Omega)$, and $|\Phi_{\delta_n}| \leq 2\delta_n \leq 2\delta$, by condition (iv), we have

$$\Big| \int_{\Omega'} a(x, Du_n) \big(\Phi_{\delta_n}(u_n) \otimes D\varphi + \Phi_{\delta_n}(u_n - u) \otimes D(w_n \varphi)\big) \, dx \Big| \leq C\delta. \tag{5.5}$$

From property (V), Hölder's inequality, and Proposition 5.3 it follows that

$$\Big| \int_{\Omega'} F_n(x, u_n) \Phi_{\delta_n}(u_n - u) w_n \varphi \, d\mu_n \Big| \leq C\delta, \tag{5.6}$$

while from property (VI) and the definition of the function Φ we get

$$\int_{\Omega'} F_n(x, u_n) \Phi_{\delta_n}(u_n) \varphi \, d\mu_n \geq 0. \tag{5.7}$$

Since $(\Phi_{\delta_n}(u_n - u) w_n \varphi)$ converges weakly to zero in $W^{1,p}_0(\Omega, \mathbf{R}^M)$ and (f_n) converges in the sense of $(\mathcal{H}_{\Omega'})$, we have

$$\langle f_n, w_n \Phi_{\delta_n}(u_n - u)\varphi \rangle = o_n. \tag{5.8}$$

Moreover, as $0 < \delta_n \leq \delta$ and the sequence (Φ_{δ_n}) is uniformly Lipschitz, it is easy to see that there exists a positive number $\tilde{\delta} \leq \delta$ such that

$$\limsup_{n \to \infty} \langle f_n, \Phi_{\delta_n}(u_n)\varphi\rangle \;=\; \langle f, \Phi_{\tilde{\delta}}(u)\varphi\rangle. \tag{5.9}$$

Finally, since $(D\Phi_{\delta_n}(u_n - u)D(u_n - u))$ converges weakly to zero in $L^p(\Omega', \mathbf{R}^M)$ we also obtain that

$$\int_{\Omega'} a(x, Du) D\Phi_{\delta_n}(u_n - u)D(u_n - u) w_n \varphi \, dx \;=\; o_n. \tag{5.10}$$

Thus, by assumptions (i)–(v), by (5.4)–(5.10), and by the definition of the function Φ, we get

$$\begin{aligned} &\int_{\{|u_n| < \delta_n\}_{\Omega'}} |Du_n|^p \varphi \, dx, + \int_{\{|u_n - u| < \delta_n\}_{\Omega'}} |D(u_n - u)|^p w_n \varphi \, dx \\ &\leq C \int_{\{\delta_n \leq |u_n - u| < 2\delta_n\}_{\Omega'}} (h + |Du| + |Du_n|)^{p-2} |D(u_n - u)|^2 \, dx \\ &+ C \int_{\{\delta_n \leq |u_n| < 2\delta_n\}_{\Omega'}} (k + \eta |Du_n|^{p-1})|Du_n| \, dx + \langle f, \Phi_{\tilde{\delta}}(u)\varphi\rangle + C\delta + o_n, \end{aligned} \tag{5.11}$$

where we also used the fact that the sequence (w_n) is bounded in $L^\infty(\Omega)$. Now, since (u_n) is bounded in $W^{1,p}(\Omega', \mathbf{R}^M)$, there exists a positive constant K such that

$$\int_{\Omega'} (h + |Du| + |Du_n|)^{p-2} |D(u_n - u)|^2 \, dx + \int_{\Omega'} (k + \eta |Du_n|^{p-1})|Du_n| \, dx \;\leq\; K.$$

In particular, if we fix $J \in \mathbf{N}$ and $\gamma > 0$, then we have

$$\begin{aligned} \sum_{j=1}^{J} \Big(&\int_{\{2^{j-1}\gamma \leq |u_n - u| < 2^j\gamma\}_{\Omega'}} (h + |Du| + |Du_n|)^{p-2} |D(u_n - u)|^2 \, dx \\ &+ \int_{\{2^{j-1}\gamma \leq |u_n| < 2^j\gamma\}_{\Omega'}} (k + \eta |Du_n|^{p-1})|Du_n| \, dx \Big) \;\leq\; K; \end{aligned}$$

so that, for every $n \in \mathbf{N}$, there exists $j(n) \in \{1, \ldots, J\}$ such that

$$\begin{aligned} &\int_{\{2^{j(n)-1}\gamma \leq |u_n - u| < 2^{j(n)}\gamma\}_{\Omega'}} (h + |Du| + |Du_n|)^{p-2} |D(u_n - u)|^2 \, dx \\ &\quad + \int_{\{2^{j(n)-1}\gamma \leq |u_n| < 2^{j(n)}\gamma\}_{\Omega'}} (k + \eta |Du_n|^{p-1})|Du_n| \, dx \;\leq\; \frac{K}{J}. \end{aligned}$$

If in (5.11) we take $\delta = 2^J\gamma$ and $\delta_n = 2^{j(n)-1}\gamma$, then we get

$$\begin{aligned} &\int_{\{|u_n| < \gamma\}_{\Omega'}} |Du_n|^p \varphi \, dx + \int_{\{|u_n - u| < \gamma\}_{\Omega'}} |D(u_n - u)|^p w_n \varphi \, dx \\ &\qquad \leq\; \frac{C}{J} + C2^J\gamma \;+\; \langle f, \Phi_{\tilde{\delta}}(u)\varphi\rangle \;+\; o_n, \end{aligned} \tag{5.12}$$

where we used the fact that $\delta_n \geq \gamma$ for every $n \in \mathbf{N}$. By Rellich's theorem the sequence (u_n) converges to u strongly in $L^p_{\text{loc}}(\Omega', \mathbf{R}^M)$, and hence, up to a subsequence, pointwise a.e. in Ω'. Thus, by Egorov's theorem, for every $\sigma > 0$ there exists a subset S of Ω', with $|S| < \sigma$, such that (u_n) converges to u uniformly on $\Omega' \setminus S$.

Now let $\varepsilon > 0$. If we choose $J \in \mathbf{N}$ and $\gamma > 0$ such that $1/J < \varepsilon$ and $\delta = 2^J\gamma = \varepsilon$, then by (5.12) we have

$$\limsup_{n\to\infty}\Big(\int_{\{|u_n|<\gamma\}_{\Omega'}} |Du_n|^p\varphi\,dx + \int_{\{|u_n-u|<\gamma\}_{\Omega'}} |D(u_n-u)|^p w_n\varphi\,dx\Big) \le C\varepsilon + \langle f, \Phi_{\tilde\delta}(u)\varphi\rangle.$$

Moreover, for n large enough we have that $\Omega' \setminus S \subseteq \{|u_n - u| < \gamma\}_{\Omega'}$ and $\{u = 0\}_{\Omega'} \setminus S \subseteq \{|u_n| < \gamma\}_{\Omega'}$, so that we get

$$\begin{aligned}\text{(5.13)}\qquad &\limsup_{n\to\infty}\Big(\int_{\{u=0\}_{\Omega'}\setminus S} |Du_n|^p\varphi\,dx + \int_{\Omega'\setminus S} |D(u_n-u)|^p w_n\varphi\,dx\Big)\\ &\quad\le C\varepsilon + \langle f, \Phi_{\tilde\delta}(u)\varphi\rangle,\end{aligned}$$

which, by using that $0 \le \tilde\delta \le \delta = \varepsilon$ and $\Phi_{\tilde\delta}(u)\varphi$ converges strongly to zero in $W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M)$ as ε tends to zero, gives

$$\text{(5.14)}\qquad \lim_{n\to\infty}\Big(\int_{\{u=0\}_{\Omega'}\setminus S} |Du_n|^p\varphi\,dx + \int_{\Omega'\setminus S} |D(u_n-u)|^p w_n\varphi\,dx\Big) = 0.$$

By the arbitrariness of σ, we get, up to a subsequence, that $(D(u_n - u)w_n)$ and $(Du_n 1_{\{|u|=0\}_{\Omega'}})$ converges to zero pointwise a.e. in Ω'. Moreover, since (w_n) converges to w strongly in $L^p(\Omega', \mathbf{R}^M)$ and by Lemma 4.5 $\{w > 0\} \supseteq \{|u| > 0\}_{\Omega'}$, this implies that (Du_n) converges pointwise to Du a.e. in $\{|u| > 0\}_{\Omega'}$ and hence, as $|Du| = 0$ a.e. in $\{|u| = 0\}_{\Omega'}$, (Du_n) converges pointwise to Du a.e. in Ω'.

Finally, since (u_n) is bounded in $W^{1,p}(\Omega', \mathbf{R}^M)$, we obtain that (u_n) converges to u strongly in $W^{1,r}(\Omega', \mathbf{R}^M)$ for every $r < p$. □

Remark 5.5. Under the same assumptions of Proposition 5.4, by (v) and Proposition 5.4 we have that $(a(x, Du_n))$ converges to $a(x, Du)$ weakly in $L^{p'}(\Omega', \mathbf{M}^{M\times N})$ and strongly in $L^r(\Omega', \mathbf{M}^{M\times N})$ for every $1 \le r < p'$. Similarly we deduce that $(a(x, D(u_n - u)))$ converges to zero weakly in $L^{p'}(\Omega', \mathbf{M}^{M\times N})$ and strongly in $L^r(\Omega', \mathbf{M}^{M\times N})$ for every $1 \le r < p'$.

6. The limit problem. In this section we shall prove the main result of this paper (Theorem 6.4). We shall consider a sequence (u_n) of solutions of problems (5.2), with $F_n \in \mathcal{F}(L)$ and $\mu_n \in \mathcal{M}_0^p(\Omega)$, which satisfies

$$\text{(6.1)}\qquad \int_{\Omega'} |Du_n|^p dx + \int_{\Omega'} |u_n|^p d\mu_n \le M,$$

where M is a positive constant which depends on the sequence (u_n). We shall show that a cluster point u of such a sequence is a solution of a variational problem similar to (5.2). Namely we shall prove that the limit problem will be of the form

$$\text{(6.2)}\qquad \begin{cases} u \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_\mu(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du)Dv\,dx + \int_{\Omega'} F(x, u)v\,d\mu = \langle f, v\rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_\mu(\Omega', \mathbf{R}^M), \end{cases}$$

where μ is a measure in $\mathcal{M}_0^p(\Omega)$ and $F(x, s)$ is a vector function in $\mathcal{F}(\alpha, C, 1/(p-1))$ for a constant C which depends only on α, β, L, N, and p.

Remark 6.1. Let $\mu \in \mathcal{M}_0^p(\Omega)$ and let $\tilde{F}, F \in \mathcal{F}(c_1, c_2, \sigma)$ be two functions such that for every $s \in \mathbf{R}^M$ $F(x,s) = \tilde{F}(x,s)$ μ-a.e. in $\{w > 0\}$, where w is the solution of problem (3.3). If in problem (6.2) we change F by $\tilde{F}$ we obtain an equivalent problem. In particular the function $F(x,s)$ can be defined arbitrarily in the set $\{w = 0\}$.

Let us introduce now a notion of convergence in the space $\mathcal{M}_0^p(\Omega) \times \mathcal{F}(c_1, c_2, \sigma)$, with $c_1 > 0$, $c_2 > 0$, and $0 < \sigma \le 1$.

Definition 6.2. *Let (μ_n) be a sequence in $\mathcal{M}_0^p(\Omega)$, let (F_n) be a sequence in $\mathcal{F}(c_1, c_2, \sigma)$, let $\mu \in \mathcal{M}_0^p(\Omega)$ and $F \in \mathcal{F}(c_1, c_2, \sigma)$. We say that the pairs (μ_n, F_n) γ^A-converge (in Ω) to the pair (μ, F) if the following property holds: for any open set $\Omega' \subseteq \Omega$, for any sequence of functionals (f_n) with $f_n \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))'$, which converges to some $f \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$ (according with Definition* 5.1*), and for any sequence (u_n) of solutions of problems* (5.2) *satisfying* (6.1)*, all cluster points of the sequence (u_n) in the weak topology of $W^{1,p}(\Omega', \mathbf{R}^M)$ satisfy problem* (6.2).

The most important property of the γ^A convergence is the following result.

Proposition 6.3. *Let $((\mu_n, F_n))$ be a sequence in $\mathcal{M}_0^p(\Omega) \times \mathcal{F}(c_1, c_2, \sigma)$ which γ^A-converges to a pair (μ, F). Then for any open set $\Omega' \subseteq \Omega$ and for any sequence (f_n), with $f_n \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))'$, which converges to some $f \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$, the unique solution u_n of the problem*

$$(6.3) \qquad \begin{cases} u_n \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du_n) Dv \, dx + \int_{\Omega'} F_n(x, u_n) v \, d\mu_n = \langle f_n, v \rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M) \end{cases}$$

converges weakly in $W_0^{1,p}(\Omega', \mathbf{R}^M)$ to the unique solution u of the problem

$$(6.4) \qquad \begin{cases} u \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du) Dv \, dx + \int_{\Omega'} F(x, u) v \, d\mu = \langle f, v \rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M). \end{cases}$$

Proof. By using u_n as a test function in (6.3) and by taking into account Lemma 5.2, we deduce that the sequence (u_n) satisfies (6.1). This implies in particular that there exists a subsequence of (u_n) which converges weakly in $W_0^{1,p}(\Omega', \mathbf{R}^M)$ to a function $u \in W_0^{1,p}(\Omega', \mathbf{R}^M)$. By the definition of γ^A-convergence, the function u satisfies (6.4). Since this problem has a unique solution, the whole sequence (u_n) converges to u. □

The following theorem gives a compactness result for the γ^A-convergence.

Theorem 6.4. *Let (μ_n) be a sequence of measures in $\mathcal{M}_0^p(\Omega)$ and let (F_n) be a sequence in $\mathcal{F}(L)$, with $L > 0$. Then there exist an increasing sequence of integers (n_j), a measure $\mu \in \mathcal{M}_0^p(\Omega)$, and a function $F \in \mathcal{F}(\alpha, C, 1/(p-1))$ such that the pairs (μ_{n_j}, F_{n_j}) γ^A-converge to (μ, F) in Ω (according to Definition* 6.2*), where C is a positive constant which depends only on α, β, L, N, and p.*

Remark 6.5. The compactness result stated in Theorem 6.4 can be proved under more general assumptions on (F_n). Namely, if the sequence (F_n) belongs to

$\mathcal{F}(c_1, c_2, \sigma)$, for some constants $c_1 > 0$, $c_2 > 0$, and $0 < \sigma \le 1$, then there exist an increasing sequence of integers (n_j), a measure $\mu \in \mathcal{M}_0^p(\Omega)$, and a function $F \in \mathcal{F}(c_1', c_2', \sigma')$ such that the pairs (μ_{n_j}, F_{n_j}) γ^A-converge to (μ, F) in Ω. The positive constants c_1 and c_2 depend only on α, β, c_1, c_2, N, p, and σ; while $\sigma' = \min\{\sigma, 1/(p-\sigma)\}$.

In order to simplify the exposition of the proof, we shall prove only the compactness result as stated in Theorem 6.4 (the proof of the general case stated in Remark 6.5 being analogous). Before proving Theorem 6.4 we need additional information on the behavior of the sequence (u_n) of solutions of problems (5.2). To this aim we shall compare (u_n) with the sequences $(w_n \psi_m)$, $\psi_m \in C_0^\infty(\Omega, \mathbf{R}^M)$, of correctors for the p-Laplacian, studied in section 4.

In Lemma 6.6 and Propositions 6.7 and 6.8, we shall consider an open set $\Omega' \subseteq \Omega$, a sequence of measures (μ_n), a sequence of functions (F_n), two sequences of functionals (f_n), (g_n), two sequences of functions (u_n), (z_n), a measure μ, two functionals f, g, and two functions u and z such that

$$\begin{cases} \mu_n, \mu \in \mathcal{M}_0^p(\Omega), \qquad F_n \in \mathcal{F}(L), \\ (\mu_n)\ \gamma^{-\Delta_p}\text{-converges to } \mu, \end{cases} \tag{6.5}$$

$$\begin{cases} f_n, g_n \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))', \\ f, g \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M))', \\ f_n \to f \ \text{ in the sense of } (\mathcal{H}_{\Omega'}), \\ g_n \to g \ \text{ in the sense of } (\mathcal{H}_{\Omega'}), \end{cases} \tag{6.6}$$

$$\begin{cases} u_n, z_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ u, z \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M), \\ u_n \rightharpoonup u \ \text{ in } W^{1,p}(\Omega', \mathbf{R}^M), \\ z_n \rightharpoonup z \ \text{ in } W^{1,p}(\Omega', \mathbf{R}^M), \end{cases} \tag{6.7}$$

$$\begin{cases} \int_{\Omega'} a(x, Du_n) Dv\, dx + \int_{\Omega'} F_n(x, u_n) v\, d\mu_n = \langle f_n, v \rangle \\ \qquad \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ \int_{\Omega'} a(x, Dz_n) Dv\, dx + \int_{\Omega'} F_n(x, z_n) v\, d\mu_n = \langle f_n, v \rangle \\ \qquad \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M). \end{cases} \tag{6.8}$$

Lemma 6.6. *Assume that* (6.5), (6.6), (6.7), *and* (6.8) *hold. Then for every function $\varphi \in C_c^\infty(\Omega')$, with $\varphi \ge 0$, we have the estimates*

$$\limsup_{n\to\infty} \Big(\int_{\Omega'} |D(u_n - u)|^p \varphi\, dx + \int_{\Omega'} |u_n|^p \varphi\, d\mu_n \Big) \le C \int_{\Omega'} |u|^p \varphi\, d\mu \tag{6.9}$$

and

$$\begin{aligned} &\limsup_{n\to\infty} \Big(\int_{\Omega'} |D((u_n - z_n) - (u - z))|^p \varphi\, dx + \int_{\Omega'} |u_n - z_n|^p \varphi\, d\mu_n \Big) \\ &\qquad \le C \Big(\int_{\Omega'} |u|^p \varphi\, d\mu + \int_{\Omega'} |z|^p \varphi\, d\mu \Big)^{\frac{p-2}{p-1}} \Big(\int_{\Omega'} |u - z|^p \varphi\, d\mu \Big)^{\frac{1}{p-1}}, \end{aligned} \tag{6.10}$$

where C is a positive constant which depends only on α, β, L, N, and p.

Proof. By Theorem 4.4 we have that u and z belong to $W^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M)$. Let $\varphi \in C_c^\infty(\Omega')$, with $\varphi \ge 0$, let w_n and w be the solutions of problems (3.5) and (3.3). By Proposition 3.6 there exists a sequence (ψ_m) in $C_0^\infty(\Omega', \mathbf{R}^M)$ such that $(w\psi_m)$ converges to $u - z$ strongly in $W^{1,p}_{\text{loc}}(\Omega') \cap L^p_\mu(\Omega')$. Thus, taking $(u_n - z_n - w_n\psi_m)\varphi$

as a test function in the difference of the equations in (6.8), we get

$$
\begin{aligned}
&\int_{\Omega'}[a(x,Du_n)-a(x,Dz_n)]D(u_n-z_n-w_n\psi_m)\varphi\,dx\\
&\quad+\int_{\Omega'}[F_n(x,u_n)-F_n(x,z_n)](u_n-z_n-w_n\psi_m)\varphi\,d\mu_n\\
&=-\int_{\Omega'}[a(x,Du_n)-a(x,Dz_n)](u_n-z_n-w_n\psi_m)\otimes D\varphi\,dx\\
&\quad+\langle f_n-g_n,(u_n-z_n-w_n\psi_m)\varphi\rangle\;=\;o_{m,n}.
\end{aligned}
\tag{6.11}
$$

Let us estimate the terms which appear in (6.11). By using assumption (ii) and Proposition 5.4, the sequences $(|a(x,Du_n)-a(x,D(u_n-u)|^{p'})$ and $(|a(x,Dz_n)-a(x,D(z_n-z)|^{p'})$ are uniformly integrable and pointwise convergent respectively to $|a(x,Du)|^{p'}$ and $|a(x,Dz)|^{p'}$. Therefore they converge strongly in $L^1(\Omega',\mathbf{M}^{M\times N})$ and hence, from (6.11), we deduce

$$
\begin{aligned}
&\int_{\Omega'}[a(x,D(u_n-u))-a(x,D(z_n-z))]D(u_n-z_n-(u-z))\varphi\,dx\\
&\quad+\int_{\Omega'}[F_n(x,u_n)-F_n(x,z_n)](u_n-z_n)\varphi\,d\mu_n\\
&=o_{m,n}-\int_{\Omega'}[a(x,Du)-a(x,Dz)]D(u_n-z_n-w_n\psi_m)\varphi\,dx\\
&+\int_{\Omega'}[a(x,D(u_n-u))-a(x,D(z_n-z))]D(w_n\psi_m-(u-z))\varphi\,dx\\
&\quad+\int_{\Omega'}[F_n(x,u_n)-F_n(x,z_n)]w_n\psi_m\varphi\,d\mu_n.
\end{aligned}
$$

Since

$$
\int_{\Omega'}[a(x,Du)-a(x,Dz)]D(u_n-z_n-w_n\psi_m)\varphi\,dx\;=\;o_{m,n}
$$

by properties (i) and (ii) of a and properties (II) and (III) of F_n, we get

$$
\begin{aligned}
&\alpha\int_{\Omega'}|D(u_n-z_n-(u-z))|^p\varphi\,dx+\alpha\int_{\Omega'}|u_n-z_n|^p\varphi\,d\mu_n\\
&\le\beta\int_{\Omega'}(h+|D(u_n-u)|+|D(z_n-z)|)^{p-2}|D(u_n-z_n-(u-z))|\,|D(w_n\psi_m-(u-z))|\varphi\,dx\\
&\quad+L\int_{\Omega'}(|u_n|+|z_n|)^{p-2}|u_n-z_n|\,|w_n\psi_m|\varphi\,d\mu_n+o_{m,n}\\
&=\beta\int_{\Omega'}(|D(u_n-u)|+|D(z_n-z)|)^{p-2}|D(u_n-z_n-(u-z))|\,|D(w_n\psi_m-(u-z))|\varphi\,dx\\
&\quad+L\int_{\Omega'}(|u_n|+|z_n|)^{p-2}|u_n-z_n|\,|w_n\psi_m|\varphi\,d\mu_n+o_{m,n}.
\end{aligned}
\tag{6.12}
$$

Using Young's inequality and then Hölder's inequality in (6.12), we obtain

$$\begin{aligned}
&\int_{\Omega'} |D((u_n - z_n) - (u - z))|^p \varphi\, dx + \int_{\Omega'} |u_n - z_n|^p \varphi\, d\mu_n \\
&\leq C\Big(\int_{\Omega'} (|D(u_n - u)| + |D(z_n - z)|)^p \varphi\, dx\Big)^{\frac{p-2}{p-1}} \Big(\int_{\Omega'} |D(w_n\psi_m - (u - z))|^p \varphi\, dx\Big)^{\frac{1}{p-1}} \\
&\quad + C\Big(\int_{\Omega'} \big(|u_n| + |z_n|\big)^p \varphi\, d\mu_n\Big)^{\frac{p-2}{p-1}} \Big(\int_{\Omega'} |w_n\psi_m|^p \varphi\, d\mu_n\Big)^{\frac{1}{p-1}} + o_{m,n}.
\end{aligned} \tag{6.13}$$

Taking $z_n = z = 0$ (and then $g_n = 0$), in estimate (6.13), by Young's inequality, we get

$$\begin{aligned}
&\int_{\Omega'} |D(u_n - u)|^p \varphi\, dx + \int_{\Omega'} |u_n|^p \varphi\, d\mu_n \\
&\quad \leq C \int_{\Omega'} |D(w_n\psi_m - u)|^p \varphi\, dx + C \int_{\Omega'} |w_n\psi_m|^p \varphi\, d\mu_n + o_{m,n},
\end{aligned}$$

which by Lemma 4.3 implies (6.9).

Finally, in order to get (6.10), it is enough to apply, in estimate (6.13), estimate (6.9) for u_n and z_n, and Lemma 4.3. □

The following proposition gives a first version of the limit problem satisfied by u.

PROPOSITION 6.7. *Let us assume* (6.5), (6.6), (6.7), *and* (6.8). *Then there exists a $w\mu$-measurable vector function H, uniquely defined μ-a.e. in Ω', such that the function u satisfies the problem*

$$\begin{cases} u \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_\mu(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du) Dv\, dx + \int_{\Omega'} Hv\, d\mu = \langle f, v\rangle \\ \forall\, v \in W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_\mu(\Omega', \mathbf{R}^M), \end{cases} \tag{6.14}$$

and

$$|H| \leq C|u|^{p-1} \qquad \mu\text{-}a.e.\ in\ \Omega'. \tag{6.15}$$

Moreover, for every $\phi \in C^\infty_c(\Omega', \mathbf{R}^M)$ we have

$$\begin{aligned}
&\int_{\Omega'} Hw\phi\, d\mu \\
&= \lim_{n\to\infty} \Big[\int_{\Omega'} a(x, D(u_n - u))\phi \otimes D(w_n - w)\, dx + \int_{\Omega'} F_n(x, u_n) w_n \phi\, d\mu_n\Big],
\end{aligned} \tag{6.16}$$

where w_n and w are the solutions of problems (3.5) *and* (3.3).

Proof. Given $\phi \in C^\infty_c(\Omega', \mathbf{R}^M)$, we take $w_n\phi$ as the test function in the equation satisfied by u_n (see (6.8)) and we get

$$\begin{aligned}
&\int_{\Omega'} a(x, Du_n)\phi \otimes Dw_n\, dx + \int_{\Omega'} a(x, Du_n) D\phi\, w_n dx \\
&\quad + \int_{\Omega'} F_n(x, u_n) w_n \phi\, d\mu_n = \langle f_n, w_n\phi\rangle.
\end{aligned} \tag{6.17}$$

By Remark 5.5 we have

$$\lim_{n\to\infty}\Big(\langle f_n, w_n\phi\rangle - \int_{\Omega'} a(x,Du_n)D\phi\, w_n dx\Big) \;=\; \langle f, w\phi\rangle - \int_{\Omega'} a(x,Du)D\phi\, w dx.$$

Let us define a distribution T in Ω' by

$$\begin{aligned}\langle T,\phi\rangle \;=\; \lim_{n\to\infty}\Big[&\int_{\Omega'} a(x,Du_n)\phi\otimes Dw_n\,dx\\ &- \int_{\Omega'} a(x,Du)\phi\otimes Dw\,dx \;+\; \int_{\Omega'} F_n(x,u_n)w_n\phi\,d\mu_n\Big]\end{aligned}$$

for every $\phi\in C_c^\infty(\Omega',\mathbf{R}^M)$. Since the norm of w_n in $W_0^{1,p}(\Omega)\cap L^p_{\mu_n}(\Omega)$ is bounded, by (6.1) and property (V) we have

$$\int_\Omega |a(x,Du_n)||Dw_n|\,dx \;+\; \int_\Omega |F_n(x,u_n)||w_n|\,d\mu_n \;\le\; C,$$

and hence T is continuous with respect to the uniform convergence and it can be represented by a vector Radon measure $(T_1,\dots,T_M)$ such that

$$\langle T,\phi\rangle \;=\; \sum_{i=1}^M\int_{\Omega'}\phi_i dT_i \qquad \forall\,\phi\in C_c^\infty(\Omega',\mathbf{R}^M), \tag{6.18}$$

where $\phi_1,\dots,\phi_M$ are the components of the vector function ϕ. Thus taking the limit in (6.17) we obtain

$$\int_{\Omega'} a(x,Du)D(w\phi)\,dx \;+\; \langle T,\phi\rangle \;=\; \langle f, w\phi\rangle. \tag{6.19}$$

Since by conditions (ii) and (iv) and Proposition 5.4, $a(x,Du_n)-a(x,D(u_n-u))$ converges to $a(x,Du)$ strongly in $L^{p'}(\Omega,\mathbf{M}^{M\times N})$, we can write

$$\langle T,\phi\rangle \;=\; \lim_{n\to\infty}\Big[\int_{\Omega'} a(x,D(u_n-u))\phi\otimes D(w_n-w)\,dx \;+\; \int_{\Omega'} F_n(x,u_n)w_n\phi\,d\mu_n\Big]. \tag{6.20}$$

Let us prove (6.16). For every $\phi\in C_c^\infty(\Omega',\mathbf{R}^M)$, with $\phi\ge 0$, by assumptions (V) and (v), Proposition 5.4, Hölder's inequality, Lemma 4.3, and estimate (6.9), we have

$$\begin{aligned}|\langle T,\phi\rangle| \;&\le\; C\limsup_{n\to\infty}\Big[\int_{\Omega'}(k(x)+\eta|D(u_n-u)|^{p-1})|D(w_n-w)||\phi|\,dx\\ &\qquad + \int_{\Omega'}|F_n(x,u_n)|w_n|\phi|\,d\mu_n\Big]\\ &\le\; C\limsup_{n\to\infty}\Big[\int_\Omega |D(u_n-u)|^{p-1}|D(w_n-w)||\phi|\,dx + \int_{\Omega'}|u_n|^{p-1}w_n|\phi|\,d\mu_n\Big]\\ &\le\; C\limsup_{n\to\infty}\Big[\Big(\int_{\Omega'}|D(u_n-u)|^p|\phi|\,dx\Big)^{\frac{p-1}{p}}\Big(\int_\Omega|D(w_n-w)|^p|\phi|\,dx\Big)^{\frac1p}\\ &\qquad +\Big(\int_{\Omega'}|u_n|^p|\phi|\,d\mu_n\Big)^{\frac{p-1}{p}}\Big(\int_{\Omega'}|w_n|^p|\phi|\,d\mu_n\Big)^{\frac1p}\Big]\\ &\le\; C\Big(\int_{\Omega'}|u|^p|\phi|\,d\mu\Big)^{\frac{p-1}{p}}\Big(\int_{\Omega'}|w|^p|\phi|\,d\mu\Big)^{\frac1p}.\end{aligned} \tag{6.21}$$

Let us denote by $|T_i|$ the total variation of the measures T_i, $i = 1\dots, M$. Taking into account that for every open subset A of Ω' we have

$$|T_i|(A) \,=\, \sup\{\langle T_i, \varphi\rangle \,:\, \varphi \in C_0^\infty(A),\ \sup|\varphi| \le 1\},$$

by (6.21) we get

$$|T_i|(A) \,\le\, C\Big(\int_A |u|^p d\mu\Big)^{\frac{p-1}{p}}\Big(\int_A |w|^p d\mu\Big)^{\frac{1}{p}} \tag{6.22}$$

for every open subset A of Ω'. Since $|u|^p\mu$, $|w|^p\mu$, and $|T_i|$ are finite measures, (6.22) holds for every Borel subset of Ω'. This implies that the measures T_i are absolutely continuous with respect to the measure $|w|^p\mu$, and hence to the measure $w\mu$. Since $w\mu$ is a σ-finite measure we can apply the Radon–Nikodým derivation theorem and we find a $w\mu$-measurable vector function $H = (H_1, \dots, H_M)$ such that

$$T_i(A) \,=\, \int_A H_i w d\mu$$

for every Borel subset A of Ω' and $i = 1, \dots, M$, so that, by (6.20) and (6.18), (6.16) holds. We can suppose that

$$H_i(x) = 0 \qquad \text{for } \mu\text{-a.e. } x \text{ in } \{w = 0\} \qquad \forall\, i = 1, \dots, M. \tag{6.23}$$

Thus by (6.22) we get

$$\int_A |H_i| w d\mu \,\le\, C\Big(\int_A |u|^p d\mu\Big)^{\frac{p-1}{p}}\Big(\int_A |w|^p d\mu\Big)^{\frac{1}{p}}$$

for every Borel subset A of Ω'. Using Young's inequality, we obtain

$$\int_A |H_i| w d\mu \,\le\, C\Big(\frac{1}{p'\varepsilon^{p'}}\int_A |u|^p d\mu + \frac{\varepsilon^p}{p}\int_A |w|^p d\mu\Big)$$

for every Borel subset A of Ω' and for every $\varepsilon > 0$. Thus (first reasoning for $\varepsilon \in \mathbf{Q}$ and then arguing by density) we get

$$|H_i(x)| w(x) \,\le\, C\Big(\frac{1}{p'\varepsilon^{p'}}|u(x)|^p + \frac{\varepsilon^p}{p}|w(x)|^p\Big) \tag{6.24}$$

for μ-a.e. x in Ω' and for every $\varepsilon > 0$. If $x \in \Omega'$ satisfies $w(x) > 0$ and (6.24) holds true for any ε, by choosing $\varepsilon = |u(x)|^{\frac{p-1}{p}}/|w(x)|^{\frac{p-1}{p}}$ in (6.24) and taking into account (6.23), we get

$$|H_i(x)| \,\le\, C|u(x)|^{p-1}, \qquad \mu\text{-a.e. } x \in \Omega',$$

and hence (6.15) is proved. Condition (6.14) follows from (6.19), (6.18), and the density result given by Proposition 3.6. Finally the vector function H is uniquely determined μ-a.e. in Ω' by (6.14) and (6.15). Indeed, by (6.14) H is uniquely determined μ-a.e. in $\{w > 0\}$, and by (6.15) we have $H = 0$ μ-a.e. in $\{|u| = 0\}_{\Omega'}$. Then the conclusion follows by Lemma 4.5. □

In order to study the dependence of the function H on the function u, let us consider a sequence of functionals (g_n) and a sequence of functions (z_n) which satisfy

(6.6), (6.7), and (6.8). By Proposition 6.7, applied to (z_n), we get that there exists a $w\mu$-measurable vector function H', uniquely defined μ-a.e. in Ω', such that

$$\text{(6.25)}\qquad \begin{cases} z \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_\mu(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Dz)Dv\,dx + \int_{\Omega'} H'v\,d\mu = \langle g, v\rangle \\ \forall\, v \in W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_\mu(\Omega', \mathbf{R}^M), \end{cases}$$

$$\text{(6.26)}\qquad |H'| \le C|z|^{p-1}, \qquad \mu\text{-a.e. in } \Omega',$$

and

$$\text{(6.27)}\qquad \begin{aligned} &\int_{\Omega'} H'w\phi\,d\mu \\ &= \lim_{n\to\infty}\Big[\int_{\Omega'} a(x, D(z_n - z))\phi \otimes D(w_n - w)\,dx + \int_{\Omega'} F_n(x, z_n)w_n\phi\,d\mu_n\Big]. \end{aligned}$$

The following proposition compares the function H with the function H'.

PROPOSITION 6.8. *The vector functions H and H' satisfy*

$$\text{(6.28)}\qquad |H - H'| \le C\big(|u| + |z|\big)^{p\frac{p-2}{p-1}}|u - z|^{\frac{1}{p-1}}, \qquad \mu\text{-a.e. in } \Omega'$$

and

$$\text{(6.29)}\qquad (H - H')(u - z) \ge \alpha|u - z|^p, \qquad \mu\text{-a.e. in } \Omega'.$$

Proof. Let us first prove (6.28). Consider $\phi \in C^\infty_c(\Omega', \mathbf{R}^M)$ and let w_n and w be the solutions of problems (3.5) and (3.3). By (6.16), (6.27), and by assumptions (ii) and (III), we have

$$\text{(6.30)}\qquad \begin{aligned} &\Big|\int_{\Omega'} (H - H')w\phi\,d\mu\Big| \\ &\le \Big|\int_{\Omega'} (a(x, D(u_n - u)) - a(x, D(z_n - z)))\phi \otimes D(w_n - w)\,dx\Big| \\ &\qquad + \Big|\int_{\Omega'} (F_n(x, u_n) - F_n(x, z_n))w_n\phi\,d\mu_n\Big| + o_n \\ &\le C\int_{\Omega'} \big(h + |D(u_n - u)| + |D(z_n - z)|\big)^{p-2}|D((u_n - z_n) - (u - z))|\,|D(w_n - w)|\,|\phi|\,dx \\ &\qquad + C\int_{\Omega'} (|u_n| + |z_n|)^{p-2}|u_n - z_n|\,|\phi|w_n\,d\mu_n + o_n \\ &\le C\int_{\Omega'} \big(|D(u_n - u)| + |D(z_n - z)|\big)^{p-2}|D((u_n - z_n) - (u - z))|\,|D(w_n - w)|\,|\phi|\,dx \\ &\qquad + C\int_{\Omega'} (|u_n| + |z_n|)^{p-2}|u_n - z_n|\,|\phi|w_n\,d\mu_n + o_n. \end{aligned}$$

By using Hölder's inequality, (6.9), applied to u_n and z_n and (6.10), we get

$$\begin{aligned} &\Big|\int_{\Omega'} (H - H')w\phi\,d\mu\Big| \\ &\le C\Big(\int_{\Omega'} |u|^p|\phi|\,d\mu + \int_{\Omega} |z|^p|\phi|\,d\mu\Big)^{\frac{p-2}{p-1}}\Big(\int_{\Omega'} |u - z|^p|\phi|\,d\mu\Big)^{\frac{1}{p(p-1)}}\Big(\int_{\Omega'} |w|^p|\phi|\,d\mu\Big)^{\frac{1}{p}}. \end{aligned}$$

Then we conclude as in the proof of Proposition 6.7 and we obtain (6.28).

In order to prove (6.29), let us consider a function $\varphi \in C_c^\infty(\Omega')$, with $\varphi \geq 0$. Using $(u_n - z_n)\varphi$ as a test function in the difference of the two equations in (6.8), we obtain

$$\begin{aligned}
&\int_{\Omega'} [a(x, Du_n) - a(x, Dz_n)] D(u_n - z_n)\varphi\, dx \\
&+ \int_{\Omega'} [a(x, Du_n) - a(x, Dz_n)](u_n - z_n) \otimes D\varphi\, dx \\
&+ \int_{\Omega'} [F_n(x, u_n) - F_n(x, z_n)](u_n - z_n)\varphi\, d\mu_n \;=\; \langle f_n - g_n, (u_n - z_n)\varphi\rangle.
\end{aligned}$$

We can rewrite this formula as

$$\begin{aligned}
&\int_{\Omega'} \big([a(x, Du_n) - a(x, Dz_n)] D(u_n - z_n) - \alpha |D(u_n - z_n)|^p\big)\varphi\, dx \\
&+ \alpha \int_{\Omega'} |D(u_n - z_n)|^p \varphi\, dx + \int_{\Omega'} [F_n(x, u_n) - F_n(x, z_n)](u_n - z_n)\varphi\, d\mu_n \\
&+ \int_{\Omega'} [a(x, Du_n) - a(x, Dz_n)](u_n - z_n) \otimes D\varphi\, dx \;=\; \langle f_n - g_n, (u_n - z_n)\varphi\rangle.
\end{aligned} \tag{6.31}$$

By assumption (II) and Theorem 4.4, we have

$$\begin{aligned}
&\alpha \int_{\Omega'} |D(u_n - z_n)|^p \varphi\, dx + \int_{\Omega'} [F_n(x, u_n) - F_n(x, z_n)](u_n - z_n)\varphi\, d\mu_n \\
&\qquad \geq\; \alpha \int_{\Omega'} |D(u - z)|^p \varphi\, dx \;+ \alpha \int_{\Omega'} |u - z|^p \varphi\, d\mu \;+\; o_n.
\end{aligned}$$

Moreover, by Remark 5.5, the sequence $(a(x, Du_n) - a(x, Dz_n))$ converges to $a(x, Du) - a(x, Dz)$ pointwise a.e. in Ω' and weakly in $L^{p'}(\Omega', \mathbf{M}^{M\times N})$. Then by condition (i) we can apply Fatou's lemma to the first integrand of (6.31) and, taking the limit, we obtain

$$\begin{aligned}
&\int_{\Omega'} \big([a(x, Du) - a(x, Dz)] D(u - z) - \alpha |D(u - z)|^p\big)\varphi\, dx \\
&\qquad + \alpha \int_{\Omega'} |D(u - z)|^p \varphi\, dx + \alpha \int_{\Omega'} |u - z|^p \varphi\, d\mu \\
&+ \int_{\Omega'} [a(x, Du) - a(x, Dz)](u - z) \otimes D\varphi\, dx \;\leq\; \langle f - g, (u - z)\varphi\rangle,
\end{aligned}$$

that is,

$$\int_{\Omega'} [a(x, Du) - a(x, Dz)] D\big(\varphi(u - z)\big)\, dx + \alpha \int_{\Omega'} |u - z|^p \varphi\, d\mu \;\leq\; \langle f - g, (u - z)\varphi\rangle.$$

Thus by (6.14) and (6.25) we get

$$\int_{\Omega'} (H - H')(u - z)\varphi\, d\mu \;\geq\; \alpha \int_{\Omega'} |u - z|^p \varphi\, d\mu$$

for every $\varphi \in C_c^\infty(\Omega')$, with $\varphi \geq 0$. This implies (6.29). □

Proposition 6.8 will imply that the function H defined by (6.16) depends on u only through its pointwise values, i.e., there exists a function $F(x, s)$ such that

$H(x) = F(x, u(x))$ μ-a.e. in Ω. This construction allows us to define the function $F(x, s)$ only on the pairs (x, s) such that $s = u(x)$, where u is the limit of a sequence of solutions of problems (5.2). We shall prove a penalization result (Theorem 6.9) which shows that, in some sense, it is possible to obtain any real number s as the "limit" of a sequence of solutions.

THEOREM 6.9. *Let $s \in \mathbf{R}^M$. For every $m \in \mathbf{N}$, let s_n^m be the unique solution of the problem*

$$\begin{cases} s_n^m \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M), \\ \displaystyle\int_\Omega a(x, Ds_n^m) Dv\, dx + \int_\Omega F_n(x, s_n^m) v\, d\mu_n \\ \displaystyle\quad = m \int_\Omega (|w_n s|^{p-2} w_n s - |s_n^m|^{p-2} s_n^m) v\, dx \\ \forall\, v \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M). \end{cases} \tag{6.32}$$

Then there exists an increasing sequence of indices (n_j) such that for every m the sequence $(s_{n_j}^m)_{j \in \mathbf{N}}$ converges to some function s^m weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$. The sequence (s^m) converges to ws strongly in $W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M)$ and satisfies

$$\lim_{m \to \infty} m \int_\Omega |s^m - ws|^p\, dx = 0.$$

Moreover, there exists a unique $w\mu$-measurable function H_s^m, with

$$|H_s^m| \le C|s^m|^{p-1}, \qquad \mu\text{-a.e. in } \Omega, \tag{6.33}$$

such that the function s^m satisfies the problem

$$\begin{cases} s^m \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu}(\Omega, \mathbf{R}^M), \\ \displaystyle\int_\Omega a(x, Ds^m) Dv\, dx + \int_\Omega H_s^m v\, d\mu \\ \displaystyle\quad = m \int_\Omega (|w_n \psi|^{p-2} w_n \psi - |s^m|^{p-2} s^m) v\, dx \\ \forall\, v \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu}(\Omega, \mathbf{R}^M). \end{cases} \tag{6.34}$$

The sequence (H_s^m) converges in $L^{p'}_\mu(\Omega, \mathbf{R}^M)$ to a function H_s which satisfies

$$|H_s| \le C|s^m|^{p-1}, \qquad \mu\text{-a.e. in } \Omega. \tag{6.35}$$

Proof. Let $s \in \mathbf{R}^M$ and let $s_n^m \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega, \mathbf{R}^M)$ be the solution of problem (6.32). Taking $(s_n^m - w_n s)$ as the test function in (6.32) we get

$$\begin{aligned}(6.36)\qquad &\int_\Omega \big(a(x, Ds_n^m) - a(x, D(w_n s))\big) D(s_n^m - w_n s)\, dx \\ &\quad + \int_\Omega \big(F(x, s_n^m) - F(x, w_n s)\big)(s_n^m - w_n s)\, d\mu_n \\ &\quad + m \int_\Omega (|s_n^m|^{p-2} s_n^m - |w_n s|^{p-2} w_n s)(s_n^m - w_n s)\, dx \\ &= -\int_\Omega a(x, D(w_n s)) D(s_n^m - w_n s)\, dx - \int_\Omega F(x, w_n s)(s_n^m - w_n s)\, d\mu_n.\end{aligned}$$

Using assumptions (i) and (iv) of a, (II) and (V) of F_n, and (4.3) we obtain

$$\begin{aligned}&\alpha \int_\Omega |D(s_n^m - w_n s)|^p dx + \alpha \int_\Omega |s_n^m - w_n s|^p d\mu_n + m 2^{2-p} \int_\Omega |s_n^m - w_n s|^p dx \\ &\le \int_\Omega \big(k(x) + \eta |D(w_n s)|^{p-1}\big) |D(s_n^m - w_n s)|\, dx + L \int_\Omega |w_n s|^{p-1} |s_n^m - w_n s| d\mu_n.\end{aligned} \tag{6.37}$$

Then, by Young's inequality and the fact that $\int_\Omega |Dw_n|^p dx + \int_\Omega |w_n|^p d\mu_n$ is bounded, it is easy to see that there exists a constant C such that

$$\int_\Omega |D(s_n^m - w_n s)|^p dx + \int_\Omega |s_n^m - w_n s|^p d\mu_n + m \int_\Omega |s_n^m - w_n s|^p dx \le C|s|^p. \tag{6.38}$$

Then there exists an increasing sequence of indices (n_j) which, by a diagonal procedure, we can assume independent on m, such that for every $m \in \mathbf{N}$ the sequence $(s_{n_j}^m)_{j\in\mathbf{N}}$ converges to some function s^m weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$. Moreover, by Theorem 4.4 we have

$$\int_\Omega |D(s^m - ws)|^p dx + \int_\Omega |s^m - ws|^p d\mu + m \int_\Omega |s^m - ws|^p dx \le C|s|^p. \tag{6.39}$$

This implies that (s^m) converges weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ to ws. In particular $|s^m - ws|$ converges to zero weakly in $W_0^{1,p}(\Omega)$, and by Theorem 3.5 we get

$$\lim_{m\to\infty} \int_\Omega |s^m - ws| w^{p-1} d\mu = \lim_{m\to\infty} \int_\Omega |s^m - ws| d\nu = 0.$$

Thus up to a subsequence (s^m) converges to ws ν-a.e. in Ω and hence by Lemma 4.5 μ-a.e. in Ω. Moreover, since by (6.39) (s^m) is bounded in $L_\mu^p(\Omega, \mathbf{R}^M)$, it converges to ws weakly in $L_\mu^p(\Omega, \mathbf{R}^M)$.

By Proposition 6.7, for every $m \in \mathbf{N}$, there exists a $w\mu$-measurable vector function H_s^m, uniquely defined μ-a.e. in Ω, which satisfies (6.33) and such that s^m is the solution of the problem (6.34). By Proposition 6.8, for every $m, k \in \mathbf{N}$, we have

$$|H_s^m - H_s^k| \le C\big(|s^m| + |s^k|\big)^{p\frac{p-2}{p-1}} |s^m - s^k|^{\frac{1}{p-1}}, \qquad \mu\text{-a.e. in } \Omega. \tag{6.40}$$

This implies that there exists a function H_s, which satisfies (6.35), such that H_s^m converges to H_s μ-a.e. in Ω. Moreover, by Proposition 6.8, for every $m, k \in \mathbf{N}$, we have

$$(H_s^m - H_s^k)(s^m - s^k) \ge \alpha |s^m - s^k|^p, \qquad \mu\text{-a.e. in } \Omega,$$

and then, taking the limit as $k \to \infty$, we obtain

$$(H_s^m - H_s)(s^m - ws) \geq \alpha |s^m - ws|^p, \qquad \mu\text{-a.e. in } \Omega. \tag{6.41}$$

Now, taking $(s^m - ws)$ as a test function in (6.34), we get

$$\begin{aligned}\int_\Omega \big(a(x, Ds^m) - a(x, D(ws))\big) D(s^m - ws)\, dx + \int_\Omega (H_s^m - H_s)(s^m - ws)\, d\mu \\ + m \int_\Omega (|s^m|^{p-2} s^m - |ws|^{p-2} ws)(s^m - ws)\, dx \\ = - \int_\Omega a(x, D(ws)) D(s^m - ws)\, dx - \int_\Omega H_s(s^m - ws)\, d\mu.\end{aligned}$$

Then by (6.41), assumption (i), and the inequality (4.3), we obtain

$$\begin{aligned}\alpha \int_\Omega |D(s^m - ws)|^p dx + \alpha \int_\Omega |s^m - ws|^p d\mu \\ \leq - \int_\Omega a(x, D(ws)) D(s^m - ws)\, dx - \int_\Omega H_s(s^m - ws)\, d\mu.\end{aligned}$$

The conclusion follows by the weak convergence of (s^m) to ws in $W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L_\mu^p(\Omega, \mathbf{R}^M)$. □

We are now in a position to prove Theorem 6.4.

Proof of Theorem 6.4. We start by defining the sequence (n_j), the measure μ, and the function F. By Theorem 3.4 we can suppose that there exists a measure $\mu \in \mathcal{M}_0^p(\Omega)$ such that the sequence (μ_n) $\gamma^{-\Delta_p}$-converges to a measure μ. This measure will be the measure which appears in the statement.

For any $q \in \mathbf{Q}^M$, let q_n^m be the solutions of the problems

$$\begin{cases} q_n^m \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L_\mu^p(\Omega, \mathbf{R}^M), \\[2ex] \displaystyle\int_\Omega a(x, Dq_n^m) Dv\, dx + \int_\Omega F_n(x, q_n^m) v\, d\mu_n \\[2ex] \displaystyle\qquad = m \int_\Omega (|w_n q|^{p-2} w_n q - |q_n^m|^{p-2} q_n^m) v\, dx \\[2ex] \forall\, v \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L_\mu^p(\Omega, \mathbf{R}^M). \end{cases} \tag{6.42}$$

By Theorem 6.9 and a diagonal argument, there exists an increasing sequence (n_j) such that for every $q \in \mathbf{Q}^M$, the sequence $(q_{n_j}^m)$ converges weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ to a function $q^m \in W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L_\mu^p(\Omega, \mathbf{R}^M)$ when j tends to infinity, and the sequence (q^m) converges strongly in $W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L_\mu^p(\Omega, \mathbf{R}^M)$ to qw when m tends to infinity. Moreover, there exists a sequence (H_q^m) in $L_\mu^{p'}(\Omega, \mathbf{R}^M)$ such that the sequence (q^m)

satisfies the problem

$$(6.43)\qquad \begin{cases} q^m \in W_0^{1,p}(\Omega,\mathbf{R}^M)\cap L^p_\mu(\Omega,\mathbf{R}^M), \\[2mm] \displaystyle\int_\Omega a(x,Dq^m)Dv\,dx + \int_\Omega H_q^m v\,d\mu \\[2mm] \displaystyle\quad = m\int_\Omega (|w_n q|^{p-2}w_n q - |q^m|^{p-2}q^m)v\,dx \\[2mm] \forall\, v\in W_0^{1,p}(\Omega,\mathbf{R}^M)\cap L^p_\mu(\Omega,\mathbf{R}^M) \end{cases}$$

and such that it converges strongly in $L^{p'}_\mu(\Omega,\mathbf{R}^M)$ to a function H_q which satisfies

$$(6.44)\qquad |H_q| \le C|wq|^{p-1}, \qquad \mu\text{-a.e. in } \Omega.$$

Applying Proposition 6.8 to q_n^m and $(q')_n^m$ and then taking the limit in H_q^m and $H_{q'}^m$, we also have

$$(6.45)\qquad \begin{aligned} &|H_q(x)-H_{q'}(x)| \le C\big(|q|+|q'|\big)^{p\frac{p-2}{p-1}}|q-q'|^{\frac{1}{p-1}}w(x)^{p-1} \\ &\quad \forall\, q,q'\in\mathbf{Q}^M,\ \mu\text{-a.e. } x \text{ in } \Omega \end{aligned}$$

and

$$(6.46)\qquad (H_q(x)-H_{q'}(x))(q-q') \ge \alpha|q-q'|^p w(x)^p \qquad \forall\, q,q'\in\mathbf{Q}^M,\ \mu\text{-a.e. } x \text{ in } \Omega.$$

We define a function $G:\Omega\times\mathbf{Q}^M\mapsto\mathbf{R}^M$ by

$$(6.47)\qquad G(x,q) = H_q(x) \qquad \forall\, q\in\mathbf{Q}^M,\ \mu\text{-a.e. } x \text{ in } \Omega$$

and then we extend G to $\Omega\times\mathbf{R}^M$ by continuity (see (6.45)). The function G satisfies

$$(6.48)\qquad \begin{cases} |G(x,s)| \le C|s|^{p-1}w(x)^{p-1}, \\ |G(x,s)-G(x,s')| \le C(|s|+|s'|)^{p\frac{p-2}{p-1}}|s_1-s_2|^{\frac{1}{p-1}}w(x)^{p-1}, \\ (G(x,s)-G(x,s'))(s-s') \ge \alpha|s-s'|^p w(x)^p \end{cases}$$

for every s and s' in $\mathbf{R}^M$ and for μ-almost every x in Ω, and it is a Carathéodory function with respect to the σ-finite measure $w\mu$. Therefore, there exists a Borel function $F:\Omega\times\mathbf{R}^M\to\mathbf{R}^M$ such that

$$(6.49)\qquad F(x,s) = G\left(x,\frac{s}{w(x)}\right)1_{\{w>0\}}(x) + \alpha|s|^{p-2}s1_{\{w=0\}} \qquad \forall\, s\in\mathbf{R}^M,\ \mu\text{-a.e. } x \text{ in } \Omega,$$

so that, by (6.48), $F\in\mathcal{F}(\alpha,C,1/(p-1))$.

In order to prove Theorem 6.4 it remains only to show that the pairs (μ_{n_j},F_{n_j}) γ^A-converge to (μ,F). To carry this out, consider an open subset Ω' of Ω and a sequence of functionals (f_{n_j}), with $f_{n_j}\in(W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_{\mu_{n_j}}(\Omega',\mathbf{R}^M))'$, which converges in the sense of $(\mathcal{H}_{\Omega'})$ to a functional $f\in(W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_\mu(\Omega',\mathbf{R}^M))'$. We have to prove that if $u_{n_j}\in W_0^{1,p}(\Omega',\mathbf{R}^M)\cap L^p_{\mu_{n_j}}(\Omega',\mathbf{R}^M)$ satisfies (5.2) and (6.1) (with n replaced by n_j), then any cluster point u of u_{n_j} in the weak topology of $W^{1,p}(\Omega', \mathbf{R}^M)$ satisfies (6.2). To simplify the notation, let us assume that the whole sequence (u_{n_j}) converges weakly in $W^{1,p}(\Omega', \mathbf{R}^M)$ to u.

By Proposition 6.7, there exists a function $H \in L^{p'}_\mu(\Omega', \mathbf{R}^M)$ such that u satisfies (6.14). Estimate (6.28), applied with u_n and z_n replaced by u_{n_j} and $q^m_{n_j}$, gives

$$|H - H^m_q| \,\le\, C\big(|u| + |q^m|\big)^{p\frac{p-2}{p-1}} |u - q^m|^{\frac{1}{p-1}}, \qquad \mu\text{-a.e. in } \Omega',$$

and therefore, taking the limit as m tends to infinity we obtain

$$|H - F(x, qw)| \,\le\, C\big(|u| + |qw|\big)^{p\frac{p-2}{p-1}} |u - qw|^{\frac{1}{p-1}}, \qquad \mu\text{-a.e. in } \Omega',$$

which implies that for any step function $\zeta = \sum_{i=i}^m q_i 1_{B_i}$, with B_i Borel subset of Ω' and q_i in $\mathbf{Q}^M$, we get

$$|H - F(x, \zeta w)| \,\le\, C\big(|u| + |\zeta w|\big)^{p\frac{p-2}{p-1}} |u - \zeta w|^{\frac{1}{p-1}}, \qquad \mu\text{-a.e. in } \Omega'.$$

Finally, Proposition 3.6 and the continuity property (III) of F imply that $H(x) = F(x, u(x))$ μ-a.e. in Ω', which concludes the proof. □

7. Corrector. In this section, we shall fix the sequence (μ_n) in $\mathcal{M}^p_0(\Omega)$ and the sequence (F_n) in $\mathcal{F}(L)$, with $L > 0$, and we shall assume that (μ_n) $\gamma^{-\Delta_p}$-converges to μ and the pairs (μ_n, F_n) γ^A-converge to (μ, F), where $\mu \in \mathcal{M}^p_0(\Omega)$ and $F \in \mathcal{F}(\alpha, C, 1/(p-1))$. This implies that in Theorem 6.9 the solutions s^m_n of the problems (6.32) converge weakly in $W^{1,p}_0(\Omega, \mathbf{R}^M)$ to s^m when n tends to infinity without extracting any subsequence. Let us define $R^m_n : \Omega \times \mathbf{R}^M \mapsto \mathbf{M}^{M\times N}$ by

$$R^m_n(x, s) = Ds^m_n - D(sw). \tag{7.1}$$

The following result gives an approach in $L^p(\Omega, \mathbf{M}^{M\times N})$ of the gradient of the solution u_n of problem (5.2).

THEOREM 7.1. *Let Ω' be an open subset of Ω. Let (u_n) be a sequence, with $u_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M)$, which converges to a function u weakly in $W^{1,p}(\Omega', \mathbf{R}^M)$ and satisfies* (6.1). *Suppose there exists a sequence (f_n), with $f_n \in (W^{1,p}_0(\Omega) \cap L^p_{\mu_n}(\Omega))'$, which converges to $f \in (W^{1,p}_0(\Omega) \cap L^p_{\mu_n}(\Omega))'$ in the sense of $(\mathcal{H}_{\Omega'})$ and such that u_n satisfies problem* (5.2).

Then, for every function $\zeta = \sum_{i=1}^l s_i 1_{K_i}$ with s_i in $\mathbf{R}^M$ and K_i closed subsets of Ω' such that $w = 0$ μ-a.e. on $K_i \cap K_j$ for $i \neq j$, we have

$$\begin{aligned} &\limsup_{m\to\infty} \lim_{n\to\infty} \int_K |Du_n - Du - R^m_n(x, \zeta)|^p dx \\ &\quad\le C\Big(\int_K |u|^p \, d\mu + \int_K |w\zeta|^p \, d\mu\Big)^{\frac{p-2}{p-1}} \Big(\int_K |u - w\zeta|^p \, d\mu\Big)^{\frac{1}{p-1}}, \end{aligned} \tag{7.2}$$

where $K = \bigcup_{i=i}^l K_i$ and C is a positive constant which depends only on α, β, and L.

Remark 7.2. The heuristic idea of Theorem 7.1 is to show that the sequence of the gradients of u_n is, except for a sequence which converges strongly to zero in $L^p(\Omega, \mathbf{R}^M)$, equal to the gradient of u plus a sequence of nonlinear functions of the variables x and $u(x)$. This explains the nonlinearity of the function F. If it were possible to apply (7.2) by replacing ζ by u/w, we would get

$$\lim_{k\to\infty} \limsup_{n\to\infty} \int_\Omega \Big|Du_n - Du - R^m_n\Big(x, \frac{u}{w}\Big)\Big|^p \, dx = 0.$$

But the choice $\zeta = u/w$ in (7.2) is not possible since we do not know, a priori, if $R_n^m(x,s)$ is a Carathéodory function; so $R_n^m(x,u(x))$ may not even be measurable. We avoid this problem using the function $w\zeta$ to approach u. This approach is always possible by Proposition 3.6 part b.

Remark 7.3. When we consider $R_n^m(x,\zeta)$, the value of ζ on $K_i \cap K_j$, $i \neq j$, is not relevant. Indeed by taking in (7.2) $u_n = u = 0$ (then $f_n = 0$) and $\zeta = s1_{K_i \cap K_j}$ we deduce

$$\limsup_{m\to\infty}\limsup_{n\to\infty}\int_{K_i\cap K_j} R_n^m(x,s)\,dx = 0 \qquad \forall\, s \in \mathbf{R}^M.$$

Remark 7.4. If K is a compact subset of Ω' such that $\mu(K) = 0$, estimate (7.2) with $\zeta = 0$ implies that Du_n converges strongly to Du in $L^p(K, \mathbf{R}^M)$.

Proof of Theorem 7.1. Let $s \in \mathbf{R}^M$ and let K be a closed subset of Ω'. By Lemma 6.6, for every $\varphi \in C_c^\infty(\Omega', \mathbf{R}^M)$, with $\varphi \geq 1_K$ in Ω', we have

$$\tag{7.3} \begin{aligned} &\limsup_{n\to\infty}\int_K |D((u_n - s_n^m) - (u - s^m))|^p\,dx \\ &\leq C\Big(\int_{\Omega'} |u|^p\varphi\,d\mu + \int_{\Omega'} |s^m|^p\varphi\,d\mu\Big)^{\frac{p-2}{p-1}}\Big(\int_{\Omega'} |u - s^m|^p\varphi\,d\mu\Big)^{\frac{1}{p-1}}. \end{aligned}$$

If φ now decreases to 1_K, by the fact that (s^m) tends to sw strongly in $W_0^{1,p}(\Omega, \mathbf{R}^M) \cap L^p_\mu(\Omega, \mathbf{R}^M)$ and from (7.3) we deduce

$$\tag{7.4} \begin{aligned} &\limsup_{m\to\infty}\limsup_{n\to\infty}\int_K |D((u_n - s_n^m) - (u - s^m))|^p\,dx \\ &\leq C\Big(\int_K |u|^p\,d\mu + \int_K |sw|^p\,d\mu\Big)^{\frac{p-2}{p-1}}\Big(\int_K |u - sw|^p\,d\mu\Big)^{\frac{1}{p-1}}. \end{aligned}$$

Moreover, by inequality

$$\big||\xi_1|^p - |\xi_2|^p\big| \leq p(|\xi_1|^{p-1} + |\xi_2|^{p-1})|\xi_1 - \xi_2| \qquad \forall\, \xi_1, \xi_2 \in \mathbf{M}^{M\times N},$$

we get

$$\begin{aligned} &\Big||D((u_n - s_n^m) - (u - s^m))|^p - |D((u_n - s_n^m) - (u - sw))|^p\Big| \\ &\leq p\Big(|D((u_n - s_n^m) - (u - s^m))|^{p-1} + |D((u_n - s_n^m) - (u - sw))|^{p-1}\Big)|D(s^m - sw)|, \end{aligned}$$

and then by the strong convergence of (s^m) in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ we deduce

$$\lim_{m\to\infty}\limsup_{n\to\infty}\int_K\Big||D((u_n - s_n^m) - (u - s^m))|^p\,dx - |D((u_n - s_n^m) - (u - sw))|^p\Big|\,dx = 0.$$

Thus by (7.4) and the definition of R_n^m, we get

$$\tag{7.5} \begin{aligned} &\limsup_{m\to\infty}\limsup_{n\to\infty}\int_K |D(u_n - u - R_n^m(x,s))|^p\,dx \\ &\leq C\Big(\int_K |u|^p\,d\mu + \int_K |sw|^p\,d\mu\Big)^{\frac{p-2}{p-1}}\Big(\int_K |u - sw|^p\,d\mu\Big)^{\frac{1}{p-1}}. \end{aligned}$$

Consider now $\zeta = \sum_{i=1}^{l} s_i 1_{K_i}$, with $s_i \in \mathbf{R}^M$ and K_i closed subsets of Ω' such that $w = 0$ μ-a.e. on $K_i \cap K_j$, for $i \neq j$. By Lemma 4.5, we also have $|u| = 0$ μ-a.e. on $K_i \cap K_j$, for $i \neq j$. Then, if $K = \bigcup_{i=1}^{l} K_i$, by using (7.5) and Hölder's inequality we get

$$\begin{aligned}
&\limsup_{m\to\infty}\limsup_{n\to\infty}\int_K |Du_n - Du - R_n^m(x,\zeta)|^p\,dx \\
&\le \limsup_{m\to\infty}\limsup_{n\to\infty}\sum_{i=1}^{l}\int_{K_i} |D(u_n - u - R_n^m(x,s_i))|^p\,dx \\
&\le C\sum_{i=1}^{l}\Big(\int_{K_i}|u|^p\,d\mu + \int_{K_i}|ws_i|^p\,d\mu\Big)^{\frac{p-2}{p-1}}\Big(\int_{K_i}|u - ws_i|^p\,d\mu\Big)^{\frac{1}{p-1}} \\
&= C\Big(\int_K |u|^p\,d\mu + \int_K |w\zeta|^p\,d\mu\Big)^{\frac{p-2}{p-1}}\Big(\int_K |u - w\zeta|^p\,d\mu\Big)^{\frac{1}{p-1}},
\end{aligned}$$

which concludes the proof. □

8. Particular cases. In this section, we shall prove that some assumptions on the function a, as homogeneity or linearity, are inherited by function F. In [6] we construct an example which shows that the function F in general can be nonlinear and nonhomogeneous.

Homogeneous case. Let a be a function which satisfies conditions (i)–(v), as at the beginning of section 5. Let us assume in addition that a satisfies the following homogeneity condition:

(vi) for a.e. $x \in \Omega$, for every $t \in \mathbf{R}$, and for every $\xi \in \mathbf{M}^{M\times N}$,

$$a(x, t\xi) = |t|^{p-2} t a(x, t\xi).$$

Moreover, let (μ_n) be a sequence in $\mathcal{M}_0^p(\Omega)$, and let (F_n) be a sequence of functions in $\mathcal{F}(L)$ which satisfies the following condition:

(VII) for every $x \in \Omega$, for every $t \in \mathbf{R}$, and for every $s \in \mathbf{R}^M$,

$$F_n(x, ts) = |t|^{p-2} t F_n(x, s).$$

Under these assumptions we have the following result.

THEOREM 8.1. *If the function a satisfies conditions* (i)–(vi) *and the sequence (F_n) satisfies conditions* (I)–(VII), *then in Theorem* 6.4 *the function F can be chosen satisfying*

$$F(x, ts) = |t|^{p-2} t F(x, s)$$

for every $x \in \Omega$, for every $t \in \mathbf{R}$, and for every $s \in \mathbf{R}^M$.

Proof. Assumptions (vi) and (VII) imply that for every $t \in \mathbf{R}$ and for every $q \in \mathbf{Q}^M$, the solution q_n^m of (6.42) satisfies

$$(tq)_n^m = tq_n^m, \qquad \mu\text{-a.e. in } \Omega,$$

where $(tq)_n^m$ is the solution of problem (6.42) with q replaced by tq, which converges, according with Theorem 6.9, to some function $(tq)^m$ weakly in $W_0^{1,p}(\Omega, \mathbf{R}^M)$ for every $m \in \mathbf{N}$. Then taking the limit as $n \to \infty$ we have

$$(tq)^m = tq^m, \qquad \mu\text{-a.e. in } \Omega.$$

Therefore, the functions H_q^m and H_{tq}^m defined by (6.43) satisfy

$$H_{tq}^m = tH_q^m, \qquad \mu\text{-a.e. in } \Omega$$

for every $t \in \mathbf{R}$ and for every $q \in \mathbf{Q}^M$. Thus, using that for every $q \in \mathbf{Q}^M$, the function $G(x,q)$ in the proof of Theorem 6.4 is defined as the limit in m of H_q^m, the continuity of $G(x,s)$ with respect to the variable s and that the function $F(x,s)$ satisfies (6.49), we conclude the proof. □

In this special case we have the following result for the correctors defined by (7.1).

THEOREM 8.2. *Assume that the function a and the sequence* (F_n) *satisfy, respectively, properties* (vi) *and* (VII). *Then, the function* R_n^m *defined by* (7.1) *satisfies*

$$R_n^m(x,ts) = tR_n^m(x,s) \tag{8.1}$$

for almost every $x \in \Omega$, *for every* $s \in \mathbf{R}^M$, *and for every* $t \in \mathbf{R}$.

Proof. Assumptions (vi) and (VII) imply that, for every $t \in \mathbf{R}$ and for every $s \in \mathbf{R}^M$, $(ts)_n^m = ts_n^m$, where s_n^m is the solution of (6.32) and $(ts)_n^m$ is the solution of problem (6.32) with s replaced by ts. Thus the conclusion follows by the definition of R_n^m. □

Linear case. Let us consider now the linear case, i.e., let us assume, with slight abuse of notation, that the function $a(x,\xi)$ is of the form $a(x)\xi$, where $a(x)$ is a measurable function from Ω on the linear applications from $\mathbf{M}^{M\times N}$ to $\mathbf{M}^{M\times N}$ which satisfies these hypotheses:

(i$_l$) there exists a constant $\alpha > 0$ such that for every $\xi \in \mathbf{M}^{M\times N}$ and for a.e. $x \in \Omega$, we have

$$a(x)\xi\xi \geq \alpha|\xi|^2;$$

(ii$_l$) there exists a constant $\beta > 0$ such that for every $\xi \in \mathbf{M}^{M\times N}$ and for a.e. $x \in \Omega$, we have

$$|a(x)\xi| \leq \beta|\xi|.$$

Remark 8.3. Hypotheses (i$_l$) and (ii$_l$) imply (i)–(v) at the beginning of section 5 for $p = 2$.

Let us denote by $\mathcal{F}_l(L)$, with $L > 0$, the class of all vector functions from $\Omega \times \mathbf{R}^M$ to $\mathbf{R}^M$ which are linear in the second argument (i.e., of the form $F(x)s$) and which satisfy the following two conditions:

(I$_l$) for every $s \in \mathbf{R}^M$ and for every $x \in \Omega$ we have

$$F(x)ss \geq \alpha|s|^2;$$

(II$_l$) for every $s \in \mathbf{R}^M$ and for every $x \in \Omega$ we have

$$|F(x)s| \leq L|s|.$$

Remark 8.4. It is easy to see that the class $\mathcal{F}_l(L)$ defined above is contained in the class $\mathcal{F}(L)$ defined in section 5.

We are now in a position to state the following result.

THEOREM 8.5. *Assume that in Theorem* 6.4, $Au = -\text{div}\,(a(x)Du)$, *with* $a(x)$ *satisfying* (i$_l$) *and* (ii$_l$), *and that the sequence* (F_n) *belongs to* $\mathcal{F}_l(L)$. *Then, the function F which appears in the statement of Theorem* 6.4 *can be chosen in the class* $\mathcal{F}_l(L')$, *with* $L' > 0$ *different, in general, from L.*

Proof. We have already shown in Theorem 8.1 that F is homogeneous in its second argument. The additivity of F can be proved essentially with the same argument. □

For the corrector result, as in section 7, let us assume that (μ_n) $\gamma^{-\Delta}$-converges to μ and that the pairs (μ_n, F_n) γ^A-converge to (μ, F) according with Definition 6.2 (where $Au = -\mathrm{div}\,(a(x)Du)$). In this case the function $R_n^m : \Omega \times \mathbf{R}^M \to \mathbf{M}^{M\times N}$ is given by $R_n^m(x,s) = Ds_n^m(x) - D(ws)(x)$, where for every $s \in \mathbf{R}^M$, s_n^m is the solution of the problem

$$\tag{8.2} \begin{cases} s_n^m \in H_0^1(\Omega, \mathbf{R}^M) \cap L^2_{\mu_n}(\Omega, \mathbf{R}^M), \\ \displaystyle\int_\Omega a(x) Ds_n^m Dv\,dx + \int_\Omega F_n(x) s_n^m v\,d\mu_n \;=\; m\int_\Omega (w_n s - s_n^m) v\,dx \\ \forall\, v \in H_0^1(\Omega, \mathbf{R}^M) \cap L^2_{\mu_n}(\Omega, \mathbf{R}^M). \end{cases}$$

Clearly, the functions R_n^m are linear in their second argument, and hence they are Carathéodory functions. This allows us to improve Theorem 7.1.

Theorem 8.6. *Let Ω' be an open subset of Ω. Let (u_n), with $u_n \in H^1(\Omega', \mathbf{R}^M) \cap L^2_{\mu_n}(\Omega', \mathbf{R}^M)$, be a sequence which converges weakly in $H^1(\Omega', \mathbf{R}^M)$ to some function u and satisfies* (6.1). *Assume also that there exists a sequence (f_n), with f_n belonging to $(H_0^1(\Omega', \mathbf{R}^M) \cap L^2_{\mu_n}(\Omega', \mathbf{R}^M))'$, converging to some functional $f \in (H_0^1(\Omega, \mathbf{R}^M) \cap L^2_\mu(\Omega, \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$, such that (u_n) satisfies the following problem:*

$$\tag{8.3} \begin{cases} u_n \in H^1(\Omega', \mathbf{R}^M) \cap L^2_{\mu_n}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_\Omega a(x) Du_n Dv\,dx + \int_\Omega F_n(x) u_n v\,d\mu_n \;=\; \langle f_n, v\rangle \\ \forall\, v \in H_0^1(\Omega', \mathbf{R}^M) \cap L^2_{\mu_n}(\Omega', \mathbf{R}^M). \end{cases}$$

Then, for every function $\psi \in H^1(\Omega', \mathbf{R}^M) \cap L^\infty(\Omega', \mathbf{R}^M)$ and for every closed set $K \subset \Omega$, we have

$$\tag{8.4} \limsup_{m\to\infty} \limsup_{n\to\infty} \int_K |Du_n - Du - R_n^m(x)\psi|^2 dx \;\le\; C\int_K |u - w\psi|^2\,d\mu.$$

In particular, if u/w belongs to $L^\infty(K, \mathbf{R}^M)$, then we have

$$\tag{8.5} \limsup_{m\to\infty} \limsup_{n\to\infty} \int_K \left| Du_n - Du - R_n^m(x)\frac{u}{w}\right|^2 dx \;=\; 0.$$

In order to prove Theorem 8.6, we need some preliminary lemmas.

Lemma 8.7. *Let $W = \sup\{\|w_n\|_{L^\infty(\Omega)}\}$. Then for every $s \in \mathbf{R}^M$, the solutions s_n^m of* (8.2) *satisfy*

$$\tag{8.6} \limsup_{m\to\infty} \limsup_{n\to\infty} \int_{\{|s_n^m| \ge 2^k W|s|\}} |D\, s_n^m|^2 dx \le C\frac{|s|^2}{k} \qquad \forall\, k \in \mathbf{N}.$$

Proof. For any $j \in \mathbf{N}$, let us consider the function $\Phi_j : \mathbf{R}^M \mapsto \mathbf{R}^M$ defined by

$$\tag{8.7} \Phi_j(\zeta) = \begin{cases} 0 & \text{if } |\zeta| \le 2^{j-1}W|s|, \\ \dfrac{|\zeta| - 2^{j-1}W|s|}{2^{j-1}W|s|}\zeta & \text{if } 2^{j-1}W|s| < |\zeta| < 2^j W|s|, \\ \zeta & \text{if } |\zeta| \ge 2^j W|s|. \end{cases}$$

Taking $\Phi_j(s_n^m)$ as a test function in (8.2), we get

$$\int_\Omega a(x)Ds_n^m D[\Phi_j(s_n^m)]\,dx + \int_\Omega F_n(x)s_n^m\Phi_j(s_n^m)\,d\mu_n + m\int_\Omega (s_n^m - w_n s)\Phi_j(s_n^m)\,dx = 0,$$

which implies

$$\begin{aligned}\alpha\int_{\{|s_n^m|\geq 2^j W|s|\}} &|D\,s_n^m|^2\,dx\\ \leq -\int_{\{2^{j-1}W|s|\leq|s_n^m|<2^jW|s|\}} &a(x)Ds_n^m D[\Phi_j(s_n^m)]\,dx + m\int_\Omega |s_n^m - w_n s||\Phi_j(s_n^m)|\,dx\\ \leq C\int_{\{2^{j-1}W|s|\leq|s_n^m|<2^jW|s|\}} &|Ds_n^m|^2\,dx + m\int_\Omega |s_n^m - w_n s||\Phi_1(s_n^m)|\,dx,\end{aligned} \tag{8.8}$$

where we used that $|\Phi_j(s_n^m)| \leq |\Phi_1(s_n^m)|$ for every $j\in\mathbf{N}$, and the fact that, in the set $\{2^{j-1}W|s| \leq |s_n^m| < 2^jW|s|\}$, we have $D[\Phi_j(s_n^m)] = Ds_n^m(2|s_n^m|-2^{j-1}W|s|)/2^{j-1}W|s|$, and hence $|D[\Phi_j(s_n^m)]| \leq 3|Ds_n^m|$.

On the other hand, by (6.38) we can see that for every $k\in\mathbf{N}$ we have

$$\sum_{j=1}^k \int_{\{2^{j-1}W|s|\leq|s_n^m|<2^jW|s|\}} |Ds_n^m|^2\,dx \;\leq\; \int_\Omega |Ds_n^m|^2\,dx \;\leq\; C|s|^2$$

and therefore, for every $k\in\mathbf{N}$, there exists $j(k)$, with $1\leq j(k)\leq k$, such that

$$\int_{\{2^{j(k)-1}W|s|\leq|s_n^m|<2^{j(k)}W|s|\}} |Ds_n^m|^2\,dx \;\leq\; C\frac{|s|^2}{k}.$$

By (8.8), applied with $j(k)$, we deduce that for any $k\in\mathbf{N}$ we have

$$\begin{aligned}&\int_{\{|s_n^m|\geq 2^kW|s|\}} |D\,s_n^m|^2\,dx\\ &\quad\leq \int_{\{|s_n^m|\geq 2^{j(k)}W|s|\}} |D\,s_n^m|^2\,dx \leq C\frac{|s|^2}{k} + m\int_\Omega |s_n^m - w_n s||\Phi_1(s_n^m)|\,dx,\end{aligned}$$

thus taking the limit as $n\to\infty$ and using that $\Phi_1(ws)=0$, we obtain

$$\begin{aligned}\limsup_{n\to\infty}\int_{\{|s_n^m|\geq 2^kW|s|\}} |D\,s_n^m|^2\,dx \;&\leq\; C\frac{|s|^2}{k} + m\int_\Omega |s^m - ws||\Phi_1(s^m)-\Phi_1(ws)|\,dx\\ &\leq\; C\frac{|s|^2}{k} + Cm\int_\Omega |s^m - ws|^2\,dx.\end{aligned}$$

Since, by Theorem 6.9, the second term on the right-hand side tends to zero when m tends to infinity, estimate (8.6) is proved. □

Lemma 8.8. *For every function $\varphi\in H_0^1(\Omega)\cap L^\infty(\Omega)$, with $\varphi\geq 0$, we have*

$$\limsup_{m\to\infty}\limsup_{n\to\infty}\int_\Omega |R_n^m(x)|^2\varphi\,dx \;\leq\; C\int_\Omega w^2\varphi\,d\mu. \tag{8.9}$$

Proof. Let $s\in\mathbf{R}^M$, with $|s|\leq 1$. Let us define $\Psi_k:\mathbf{R}^M\mapsto\mathbf{R}^M$ by $\Psi_k(\zeta) = \zeta - \Phi_k(\zeta)$, where Φ_k is the function defined by (8.7). Taking $\Psi_k(s_n^m - w_n s)\varphi$, with

$\varphi \in H_0^1(\Omega) \cap L^\infty(\Omega)$ and $\varphi \geq 0$, as a test function in (8.2), we have

$$\begin{aligned}&\int_\Omega a(x) Ds_n^m D[\Psi_k(s_n^m - w_n s)]\varphi\, dx + \int_\Omega a(x) Ds_n^m \Psi_k(s_n^m - w_n s) \otimes D\varphi\, dx\\&+ m\int_\Omega (s_n^m - w_n s)\Psi_k(s_n^m - w_n s)\varphi\, dx + \int_\Omega F_n(x) s_n^m \Psi_k(s_n^m - w_n s)\varphi\, d\mu_n = 0,\end{aligned}$$

where the second term tends to zero when n and then m tend to infinity and where the third term in the left-hand side is positive. This permits us to write

$$\begin{aligned}&\alpha\int_{\{|s_n^m - w_n s| \leq 2^{k-1}W|s|\}} |D(s_n^m - w_n s)|^2\varphi\, dx + \int_\Omega F_n(x)(s_n^m - w_n s)\Psi_k(s_n^m - w_n s)\varphi\, d\mu_n\\&\qquad \leq C\int_{\{2^{k-1}W|s| < |s_n^m - w_n s| < 2^k W|s|\}} |Ds_n^m||D(s_n^m - w_n s)|\varphi\, dx\\&\qquad + C\int_{\{|s_n^m - w_n s| \leq 2^{k-1}W|s|\}} |D(w_n s - ws)||D(s_n^m - w_n s)|\varphi\, dx\\&\qquad + C\int_{\{|s_n^m - w_n s| \leq 2^{k-1}W|s|\}} |Dws||D(s_n^m - w_n s)|\varphi\, dx\\&\qquad + C\int_\Omega |w_n s||\Psi_k(s_n^m - w_n s)|\varphi\, d\mu_n + o_{m,n}.\end{aligned} \tag{8.10}$$

Since for $k \geq 2$

$$|s_n^m| \geq |s_n^m - w_n s| - |w_n s| \geq |s_n^m - w_n s| - W|s| \geq 2^{k-2}W|s| \tag{8.11}$$

in the set $\{|s_n^m - w_n s| \geq 2^{k-1}W|s|\}$, by Hölder's inequality, (6.38), and Lemma 8.7, we obtain

$$\int_{\{2^{k-1}W|s| < |s_n^m - w_n s| < 2^k W|s|\}} |Ds_n^m||D(s_n^m - w_n s)|\varphi\, dx \leq C\frac{\|\varphi\|_{L^\infty(\Omega)}}{\sqrt{k-2}}|s|^2 + o_{m,n}.$$

By the definition of Ψ_k and (I_l), we have

$$\begin{aligned}\int_\Omega F_n(x)(s_n^m - w_n s)\Psi_k(s_n^m - w_n s)\varphi\, d\mu_n &\geq \alpha\int_\Omega (s_n^m - w_n s)\Psi_k(s_n^m - w_n s)\varphi\, d\mu_n\\&\geq \alpha\int_\Omega |\Psi_k(s_n^m - w_n s)|^2\varphi\, d\mu_n.\end{aligned}$$

Therefore, using Young's inequality in (8.10) and taking into account that $|s| \leq 1$ and that the third term of the right-hand side of (8.10) tends to zero when n and m tend to infinity, we get

$$\begin{aligned}&\int_{\{|s_n^m - w_n s| \leq 2^{k-1}W|s|\}} |D(s_n^m - w_n s)|^2\varphi\, dx + \int_\Omega |\Psi_k(s_n^m - w_n s)|^2\varphi\, d\mu_n\\&\leq C\left[\frac{\|\varphi\|_{L^\infty(\Omega)}}{\sqrt{k-2}} + \int_\Omega |D(w_n s - ws)|^2\varphi\, dx + \int_\Omega |w_n s|^2\varphi\, d\mu_n\right] + o_{m,n}\\&\qquad \leq C\frac{\|\varphi\|_{L^\infty(\Omega)}}{\sqrt{k-2}} + C\int_\Omega w^2\varphi\, d\mu + o_{m,n},\end{aligned} \tag{8.12}$$

where in the last inequality we used Lemma 4.2. Thus by (8.12), (8.11), and Lemmas 8.7 and 4.2, we have

$$\int_\Omega |R_n^m(x)s|^2\varphi\,dx \;=\; \int_\Omega |D(s_n^m - ws)|^2\varphi\,dx$$

$$\le\; 2\int_{\{|s_n^m-w_ns|\le 2^{k-1}W|s|\}} |D(s_n^m-w_ns)|^2\varphi\,dx+2\int_{\{|s_n^m-w_ns|\le 2^{k-1}W|s|\}} |D(w_ns-ws)|^2\varphi\,dx$$

$$+2\int_{\{|s_n^m-w_ns|> 2^{k-1}W|s|\}} |Ds_n^m|^2\varphi\,dx+2\int_{\{|s_n^m-w_ns|> 2^{k-1}W|s|\}} |D(ws)|^2\varphi\,dx$$

$$\le\; C\|\varphi\|_{L^\infty(\Omega)}\Big(\frac{1}{\sqrt{k-2}}+\frac{1}{k-2}\Big)+C\int_\Omega w^2\varphi\,d\mu+o_{m,n},$$

which by the arbitrariness of k implies

$$\limsup_{m\to\infty}\limsup_{n\to\infty}\int_\Omega |R_n^m(x)s|^2\varphi\,dx \;\le\; C\int_\Omega w^2\varphi\,d\mu.$$

Since

$$|R_n^m(x)| = \max\{|R_n^m(x)s| : |s|\le 1\} \le \sum_{i=1}^N |R_n^m(x)e_i|,$$

where $\{e_i : i \le i \le N\}$ is the canonical basis of $\mathbf{R}^N$, Lemma 8.8 is proved. □

Remark 8.9. If in Lemma 8.8, φ belongs to $C_c^\infty(\Omega)$, then estimate (8.9) may be easily deduced from estimate (6.9) in Lemma 6.6. Remark also that Lemmas 8.7 and 8.8 can be easily generalized to the nonlinear case.

Proof of Theorem 8.6. By Lemma 8.8, for every closed $K \subset \Omega'$ and for every function $\psi \in H^1(\Omega', \mathbf{R}^M)$, with $\psi \ge 0$, we have

$$\limsup_{m\to\infty}\limsup_{n\to\infty}\int_K |R_m^n(x)|^2\psi\,dx \;\le\; C\int_K w^2\psi\,d\mu. \tag{8.13}$$

Indeed it is enough in (8.9) to take φ equals to $\varphi_n\psi$, with $\varphi_n \in H_0^1(\Omega', \mathbf{R}^M) \cap L^\infty(\Omega', \mathbf{R}^M)$ decreasing to the characteristic function of K.

Consider $\psi \in H^1(\Omega', \mathbf{R}^M) \cap L^\infty(\Omega', \mathbf{R}^M)$ and let K be a closed subset of Ω'. By Theorem 7.1, for any function $\zeta = \sum_{i=1}^l s_i 1_{K_i}$, with $s_i \in \mathbf{R}$ and K_i closed subsets of Ω', such that $K = \bigcup_{i=1}^l K_i$ and $w = 0$ μ-a.e. on $K_i \cap K_j$, for $i \ne j$, we have

$$\int_K |Du_n - Du - R_n^m(x)\psi|^2\,dx$$

$$\le\; 2\int_K |Du_n - Du - R_n^m(x)\zeta|^2\,dx + 2\int_K |R_n^m(x)|^2|\psi-\zeta|^2\,dx$$

$$\le\; C\int_K |u-w\zeta|^2\,d\mu + 2\sum_{i=1}^l\int_{K_i} |R_n^m(x)|^2|\psi - s_i|^2\,dx + o_{m,n} \tag{8.14}$$

$$\le\; C\int_K |u-w\zeta|^2\,d\mu + C\sum_{i=1}^l\int_{K_i} |w\psi - s_iw|^2\,d\mu + o_{m,n}$$

$$=\; C\int_K |u-w\zeta|^2\,d\mu + C\int_K |w\psi - w\zeta|^2\,d\mu + o_{m,n},$$

where we used (8.13). In order to obtain (8.4), it is enough to take $\zeta = \zeta_k$, where (ζ_k) is a sequence of step functions such that $(w\zeta_k)$ converges strongly to $w\psi$ in $L^2_\mu(\Omega', \mathbf{R}^M)$.

Assume now that u/w belongs to $L^\infty(K, \mathbf{R}^M)$ and take $\varepsilon > 0$. By estimates (8.4) and (8.13), we get

$$\begin{aligned}
&\int_K \left| Du_n - Du - R_n^m(x)\frac{u}{w}\right|^2 dx \\
&\le \int_K \left| Du_n - Du - R_n^m(x)\frac{u}{w+\varepsilon}\right|^2 dx + \int_K |R_n^m(x)|^2 \left|\frac{u}{w+\varepsilon} - \frac{u}{w}\right|^2 dx \\
&\le \int_K \left| u - \frac{wu}{w+\varepsilon}\right|^2 d\mu + C\varepsilon^2 \left\|\frac{u}{w}\right\|^2_{L^\infty(K,\mathbf{R}^M)} \int_K |R_n^m(x)|^2 \frac{1}{(w+\varepsilon)^2}\, d\mu + o_{m,n} \\
&\le \int_K \left| u - \frac{wu}{w+\varepsilon}\right|^2 d\mu + C\varepsilon^2 \left\|\frac{u}{w}\right\|^2_{L^\infty(K,\mathbf{R}^M)} \int_K \frac{w^2}{(w+\varepsilon)^2}\, d\mu + o_{m,n}.
\end{aligned} \tag{8.15}$$

By using that u belongs to $L^2_\mu(\Omega, \mathbf{R}^M)$ and the dominated convergence theorem, the first integral of the right-hand side of (8.15) tends to zero when ε tends to zero. Since

$$\frac{\varepsilon w}{(w+\varepsilon)^2} = \frac{\varepsilon}{w+\varepsilon}\frac{w}{w+\varepsilon} \le 1,$$

and hence by the fact that $\nu = w^{p-1}\mu$ is a Radon measure,

$$\varepsilon \int_K \frac{w^2}{(w+\varepsilon)^2}\, d\mu \le \int_K w\, d\mu < +\infty,$$

we get that the second integral on the right-hand side of (8.15) tends to zero when ε tends to zero. We deduce (8.5) taking the limit in n, m, and then in ε. □

9. Asymptotically equivalent operators. We saw in the previous sections that the properties of the function F which appears in the limit problem (6.2) are strictly related to the properties of the function a which define the differential operator A. The next proposition shows, in some sense, how the function F depends on the behavior of $a(x, \xi)$ when $|\xi|$ is large.

Let $\tilde{a} : \Omega \times \mathbf{M}^{M\times N} \to \mathbf{M}^{M\times N}$ be a Carathéodory function which satisfies conditions (i)–(v), and suppose that the following property

$$\lim_{|\xi|\to\infty} \frac{|a(x,\xi) - \tilde{a}(x,\xi)|}{|\xi|^{p-1}} = 0 \tag{9.1}$$

holds uniformly with respect to x in Ω. Let $\tilde{A}$ be the differential operator given by $\tilde{A}u = -\mathrm{div}\,(\tilde{a}(x, Du))$.

PROPOSITION 9.1. *Suppose that the pair* (μ_n, F_n)*, according to Definition* 6.2*,* γ^A*-converges to* (μ, F)*.*

If the functions a and $\tilde{a}$ satisfy condition (9.1)*, then we also have that* (μ_n, F_n) $\gamma^{\tilde{A}}$*-converges to* (μ, F)*.*

Proof. According to the definition of the $\gamma^{\tilde{A}}$-convergence, we have to show that for any open subset Ω' of Ω, for any sequence of functionals (f_n), with $f_n \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))'$, which converges to some $f \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap$

$L^p_\mu(\Omega', \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$, and for any sequence (u_n) which satisfies (6.1) and

$$\begin{cases} u_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} \tilde{a}(x, Du_n) Dv\, dx + \int_{\Omega'} F_n(x, u_n) v\, d\mu_n \;=\; \langle f_n, v\rangle \\ \forall\, v \in W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \end{cases} \tag{9.2}$$

every cluster point of the sequence (u_n) in the weak topology of $W^{1,p}(\Omega', \mathbf{R}^M)$ satisfies problem

$$\begin{cases} u \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} \tilde{a}(x, Du) Dv\, dx + \int_{\Omega'} F(x, u) v\, d\mu \;=\; \langle f, v\rangle \\ \forall\, v \in W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M). \end{cases} \tag{9.3}$$

If u_n satisfies (9.2), then it also satisfies

$$\begin{cases} u_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du_n) Dv\, dx + \int_{\Omega'} F_n(x, u_n) v\, d\mu_n \;=\; \langle g_n, v\rangle \\ \forall\, v \in W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \end{cases}$$

where $g_n = f_n - \mathrm{div}\,[a(x, Du_n) - \tilde{a}(x, Du_n)]$. Therefore, once we show that $(\mathrm{div}\,[a(x, Du_n) - \tilde{a}(x, Du_n)])$ converges in the sense of $(\mathcal{H}_{\Omega'})$ to $\mathrm{div}\,[a(x, Du) - \tilde{a}(x, Du)]$, by the γ^A-convergence of (μ_n, F_n) to (μ, F), we can deduce that u satisfies (9.3).

In order to prove that $(-\mathrm{div}\,[\tilde{a}(x, Du_n) - a(x, Du_n)])$ converges in the sense of $(\mathcal{H}_{\Omega'})$, let us consider $v_n \in W^{1,p}_0(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M)$ such that (v_n) converges weakly to some v in $W^{1,p}_0(\Omega', \mathbf{R}^M)$. Since by Proposition 5.4 the sequence (Du_n) converges to Du pointwise a.e. in Ω', by Egorov's theorem, for every $\delta > 0$, there exists a set $E \subseteq \Omega'$, with $|E| < \delta$, such that (Du_n) converges uniformly to Du in $\Omega' \setminus E$. Thus we get

$$\begin{aligned} &\lim_{n\to\infty} \int_{\Omega'} [\tilde{a}(x, Du_n) - a(x, Du_n)] Dv_n\, dx \\ &= \int_{\Omega' \setminus E} [\tilde{a}(x, Du) - a(x, Du)] Dv\, dx + \lim_{n\to\infty} \int_E [\tilde{a}(x, Du_n) - a(x, Du_n)] Dv_n\, dx. \end{aligned} \tag{9.4}$$

Let us estimate the last limit in (9.4). By (9.1), for every $\varepsilon > 0$ there exists $M > 0$ such that

$$|\tilde{a}(x, \xi) - a(x, \xi)| \;\le\; \varepsilon |\xi|^{p-1}$$

whenever $|\xi| > M$. Thus, since (v_n) and (u_n) are bounded in $W^{1,p}(\Omega', \mathbf{R}^M)$, by

Hölder's inequality, we get

$$\begin{aligned}
&\lim_{n\to\infty}\Big|\int_E [\tilde a(x,Du_n)-a(x,Du_n)]Dv_n\,dx\Big| \\
&\qquad\le \lim_{n\to\infty}\int_{E\cap\{|Du_n|>M\}_{\Omega'}} |\tilde a(x,Du_n)-a(x,Du_n)||Dv_n|\,dx \qquad (9.5)\\
&\quad+\int_{E\cap\{|Du_n|\le M\}_{\Omega'}} |\tilde a(x,Du_n)-a(x,Du_n)||Dv_n|\,dx \;\le\; C(\varepsilon+M^{p-1}\delta^{(p-1)/p}).
\end{aligned}$$

Now taking the limit as δ goes to zero and then the limit as M goes to infinity, by (9.4) and (9.5) we obtain

$$\lim_{n\to\infty}\int_{\Omega'}[\tilde a(x,Du_n)-a(x,Du_n)]Dv_n\,dx \;=\; \int_{\Omega'}[\tilde a(x,Du)-a(x,Du)]Dv\,dx,$$

which concludes the proof. □

Corollary 9.2. *Let (F_n) be a sequence in $\mathcal{F}(L)$ which satisfies condition* (VI) *and assume that the function a satisfies the following condition: There exists a Carathéodory function $\tilde a$ such that*

$$\lim_{t\to\infty}\frac{a(x,t\xi)}{|t|^{p-2}t} \;=\; \tilde a(x,\xi) \qquad (9.6)$$

uniformly in x, for every $\xi\in\mathbf{M}^{M\times N}$.

Suppose that the pair (μ_n,F_n), according to Definition 6.2, *γ^A-converges to (μ,F). Then the function F also satisfies condition* (VI).

Proof. It is easy to see that $\tilde a$ satisfies conditions (i)–(vi) and that condition (9.6) implies condition (9.1). Thus by the previous theorem the sequence of pairs (F_n,μ_n) $\gamma^{\tilde A}$-converges to (μ,F) and by Theorem 8.1 the function F satisfies condition (VI). □

10. General operators. In this section we shall prove that the results given in the previous sections hold for a class of more general operators. Actually, let $2\le p<+\infty$ and let $b:\Omega\times\mathbf{R}^M\times\mathbf{M}^{M\times N}\mapsto\mathbf{M}^{M\times N}$ be a Carathéodory function such that:

(i′) there exists a constant $\alpha>0$ such that

$$(b(x,0,\xi_1)-b(x,0,\xi_2))(\xi_1-\xi_2)\ge\alpha|\xi_1-\xi_2|^p$$

for every $s\in\mathbf{R}^M$, for every $\xi_1,\xi_2\in\mathbf{M}^{M\times N}$, and for a.e. $x\in\Omega$;

(ii′) there exists a constant $\beta>0$ and a function $h\in L^{\frac{p}{p-2}}(\Omega)$ $(p/(p-2)=+\infty$ if $p=2)$ such that

$$|b(x,0,\xi_1)-b(x,0,\xi_2)| \;\le\; \beta(h(x)+(|\xi_1|+|\xi_2|)^{p-2})|\xi_1-\xi_2|$$

for every $\xi_1,\xi_2\in\mathbf{M}^{M\times N}$ and for a.e. $x\in\Omega$;

(iii′) there exists a constant $\gamma>0$ and a function $k\in L^{p'}(\Omega)$ such that

$$|b(x,s_1,\xi)-b(x,s_2,\xi)| \;\le\; \gamma\big(k(x)+(|s_1|+|s_2|)^q+|\xi|^r\big)\min\{|s_1-s_2|,1\}$$

for every $s_1,s_2\in\mathbf{R}^M$, for every $\xi\in\mathbf{M}^{M\times N}$ and for a.e. $x\in\Omega$, where q and r are constants which satisfy $0\le q<N(p-1)/(N-p)$ if $p<N$, $q\ge0$ if $p\ge N$ and $0\le r<p-1$.

(iv′) $b(\cdot, 0, 0) \in L^{p'}(\Omega)$.

Under these hypotheses on the operator $Bu = -\mathrm{div}(b(x, u, Du))$, we have the following generalizations of Definition 6.2 and Theorem 6.4.

Definition 10.1. Let (μ_n) be a sequence in $\mathcal{M}_0^p(\Omega)$, let (F_n) be a sequence in $\mathcal{F}(c_1, c_2, \sigma)$, let $\mu \in \mathcal{M}_0^p(\Omega)$ and $F \in \mathcal{F}(c_1, c_2, \sigma)$. We say that the pairs (μ_n, F_n) γ^B-*converge (in* Ω*)* to the pair (μ, F) if the following property holds: for any open set $\Omega' \subseteq \Omega$, for any sequence of functionals (f_n), with $f_n \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M))'$, which converges to some $f \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$ (according to Definition 5.1), and for any sequence (u_n) of solutions of the problems

$$\tag{10.1} \begin{cases} u_n \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} b(x, u_n, Du_n) Dv\, dx + \int_{\Omega'} F_n(x, u_n) v\, d\mu_n \,=\, \langle f_n, v\rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_n}(\Omega', \mathbf{R}^M) \end{cases}$$

satisfying (6.1), all cluster points of the sequence (u_n) in the weak topology of $W^{1,p}(\Omega', \mathbf{R}^M)$ satisfy the following problem:

$$\tag{10.2} \begin{cases} u \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} b(x, u, Du) Dv\, dx \,+ \int_{\Omega'} F(x, u) v\, d\mu \,=\, \langle f, v\rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M). \end{cases}$$

Remark 10.2. If (u_n) is a sequence of solutions of problems (10.1) then the assertion of Proposition 5.4 can be proved using the same argument.

Theorem 10.3. *Let (μ_n) be a sequence of measures in $\mathcal{M}_0^p(\Omega)$ and let (F_n) be a sequence in $\mathcal{F}(L)$, with $L > 0$. Then there exists an increasing sequence of integers (n_j), a measure $\mu \in \mathcal{M}_0^p(\Omega)$, and a function $F \in \mathcal{F}(\alpha, C, 1/(p-1))$ such that the pairs (μ_{n_j}, F_{n_j}) γ^B-converge to (μ, F) in Ω (according to Definition* 10.1*).*

Proof. The above hypotheses on $b(x, s, \xi)$ imply that the application $a : \Omega \times \mathbf{M}^{M\times N} \mapsto \mathbf{M}^{M\times N}$ defined by $a(x, \xi) = b(x, 0, \xi) - b(x, 0, 0)$ satisfies conditions (i)–(v) in section 5 and then, by Theorem 6.4, there exists an increasing sequence of integers (n_j), a measure $\mu \in \mathcal{M}_0^p(\Omega)$, and a function $F \in \mathcal{F}(\alpha, C, 1/(p-1))$ such that the pairs (μ_{n_j}, F_{n_j}) γ^A-converge to (μ, F) in Ω (according to Definition 6.2). Let us see that the pairs (μ_{n_j}, F_{n_j}) γ^B-converge to (μ, F) in Ω (according to Definition 10.1). Let us consider a sequence of functionals (f_{n_j}), with $f_{n_j} \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_{n_j}}(\Omega', \mathbf{R}^M))'$, which converges to some $f \in (W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu}(\Omega', \mathbf{R}^M))'$ in the sense of $(\mathcal{H}_{\Omega'})$, a sequence (u_{n_j}) which satisfies (10.1) (with n replaced by n_j) and (6.1), and a cluster point u of the sequence (u_{n_j}) in the weak topology of $W^{1,p}(\Omega', \mathbf{R}^M)$. We have to prove that u satisfies problem (10.2). In order to simplify the notation, we shall still denote by (u_{n_j}) the subsequence of (u_{n_j}) which converges weakly in $W^{1,p}(\Omega', \mathbf{R}^M)$ to u. By (10.1), the sequence (u_{n_j}) satisfies

$$\tag{10.3} \begin{cases} u_{n_j} \in W^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_{n_j}}(\Omega', \mathbf{R}^M), \\ \displaystyle\int_{\Omega'} a(x, Du_{n_j}) Dv\, dx \,+ \int_{\Omega'} F(x, u_{n_j}) v\, d\mu \,=\, \langle f_{n_j}, v\rangle \,-\, \langle g_{n_j}, v\rangle \\ \forall\, v \in W_0^{1,p}(\Omega', \mathbf{R}^M) \cap L^p_{\mu_{n_j}}(\Omega', \mathbf{R}^M), \end{cases}$$

where $g_{n_j} = -\mathrm{div}\big(b(x, u_{n_j}, Du_{n_j}) - b(x, 0, Du_{n_j}) - b(x, 0, 0)\big)$. To conclude the proof it is enough to show that the sequence (g_{n_j}) converges in the sense of $(\mathcal{H}_{\Omega'})$ to the functional $g = -\mathrm{div}\big(b(x, u, Du) - b(x, 0, Du) - b(x, 0, 0)\big)$. By (iii$'$) we have

$$\big|b(x, u_{n_j}, Du_{n_j}) - b(x, 0, Du_{n_j})\big| \leq \gamma(k + |u_{n_j}|^q + |Du_{n_j}|^r)|u_{n_j}|. \tag{10.4}$$

By Remark 10.2, Du_{n_j} converges pointwise to Du, and then the left-hand side of (10.4) converges pointwise to $b(x, u, Du) - b(x, 0, Du)$, and the power p' of the right-hand side is uniformly integrable. This implies that $(b(x, u_{n_j}, Du_{n_j}) - b(x, 0, Du_{n_j}))$ converges strongly in $L^{p'}(\Omega')$ to $b(x, u, Du) - b(x, 0, Du)$, which concludes the proof. □

REFERENCES

[1] L. Boccardo and F. Murat, *Almost everywhere convergence of the gradients of solutions to elliptic and parabolic equations*, Nonlinear Anal., 19 (1992), pp. 581–597.

[2] G. Buttazzo and G. Dal Maso, *Shape optimization for Dirichlet problems: Relaxed solutions and optimality conditions*, Appl. Math. Optim., 23 (1991), pp. 17–49.

[3] J. Casado-Diaz, *Homogenization of Dirichlet problems for monotone operators in varying domains*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 457–478.

[4] J. Casado-Diaz, *Existence of a sequence satisfying Cioranescu-Murat conditions in homogenization of Dirichlet problems in perforated domains*, Rend. Mat. Appl. (7), 16 (1996), pp. 387–413.

[5] J. Casado-Diaz, *Homogenization of pseudomonotone Dirichlet problems in varying domains*, J. Math. Pures Appl., to appear.

[6] J. Casado-Diaz and A. Garroni, *A non homogeneous extra term for the limit of Dirichlet problems in perforated domains*, in Homogenization and Applications to Material Sciences, Math. Sciences and Appl. Series, Gakkokotosho, 1995.

[7] D. Cioranescu and F. Murat, *Un Term Étrange Venu D'Ailleurs*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France seminar, Vol. II and III, H. Brézis and J.-L. Lions, eds., Research Notes in Math. 60 and 70, Pitman, London, 1982, pp. 98–138 and pp. 154–178.

[8] G. Dal Maso, *On the integral representation of certain local functionals*, Ricerche Mat., 32 (1983), pp. 85–113.

[9] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Boston, MA, 1993.

[10] G. Dal Maso and A. Defranceschi, *Limits of nonlinear Dirichlet problems in varying domains*, Manuscripta Math., 61 (1988), pp. 251–278.

[11] G. Dal Maso and A. Garroni, *New results on the asymptotic behaviour of Dirichlet problems in perforated domains*, Math. Models Methods Appl. Sci., 3 (1994), pp. 373–407.

[12] G. Dal Maso, A. Garroni, and I. V. Skrypnik, *A capacitary method for the asymptotic analysis of Dirichlet problems for monotone operators*, J. Anal. Math., 71 (1997), pp. 263–313.

[13] G. Dal Maso and U. Mosco, *Wiener-criterion and Γ-convergence*, Appl. Math. Optim., 15 (1987), pp. 15–63.

[14] G. Dal Maso and U. Mosco, *Wiener-criteria and energy decay for relaxed Dirichlet problems*, Arch. Rational Mech. Anal., 95 (1986), pp. 345–387.

[15] G. Dal Maso and F. Murat, *Asymptotic behaviour and correctors for Dirichlet problems in perforated domains with homogeneous monotone operators*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 239–290.

[16] G. Dal Maso and F. Murat, *Dirichlet problems in perforated domains for homogeneous monotone operators on H_0^1*, in Calculus of Variations, Homogenization and Continuum Mechanics, G. Bouchitté, G. Buttazzo, and P. Suquet, eds., Series Adv. Math. Appl. Sci. 18, World Scientific, Singapore, 1994, pp. 177–202.

[17] G. Dal Maso and I. V. Skrypnik, *Capacitary theory for monotone operators*, Potential Anal., 7 (1997), pp. 765–803.

[18] A. Defranceschi and E. Vitali, *Limits of minimum problems with convex obstacles for vector valued functions*, Appl. Anal., 52 (1994), pp. 1–33.

[19] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[20] E. Ya. Khruslov, *The method of orthogonal projections and the Dirichlet problems in domains with a fine-grained boundary*, Math. USSR-Sb., 17 (1972), pp. 37–59.

[21] N. Labani and C. Picard, *Homogenization of a nonlinear Dirichlet problem in a periodically perforated domain*, in Recent Advances in Nonlinear Elliptic and Parabolic Problems, P. Bénilan, M. Chipot, L. C. Evans, and M. Pierre, eds., Pitman Res. Notes Math. Ser. 208, Longman, Harlow, 1989, pp. 294–305.

[22] A. V. Marchenko and E. Ya. Khruslov, *Boundary Value Problems in Domains with Fine-Granulated Boundaries*, Naukova Dumka, Kiev, 1974 (in Russian).

[23] J. L. Lions, *Quelques Méthodes de résolution des Problémes aux Limites Non Linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[24] I. V. Skrypnik, *Nonlinear Elliptic Boundary Value Problems*, Teubner-Verlag, Leipzig, 1986.

[25] I. V. Skrypnik, *Asymptotic behaviour of solutions of nonlinear elliptic problems in perforated domains*, Math. Sbornik, 184, 10 (1993), pp. 67–90.

[26] I. V. Skrypnik, *Homogenization of nonlinear Dirichlet problems in perforated domains of general structure*, Mat. Sb. (N.S.), to appear.

[27] I. V. Skrypnik, *New conditions for the homogenization of nonlinear Dirichlet problems in perforated domains*, Ukraïn. Mat. Zh., 48 (1996), pp. 675–694.

[28] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. Math. Anal.
Vol. 31, No. 3, pp. 625–650

SCATTERING THEORY AND SELF-SIMILAR SOLUTIONS FOR THE NONLINEAR SCHRÖDINGER EQUATION*

THIERRY CAZENAVE[†] AND FRED B. WEISSLER[‡]

Abstract. We study a certain class of mild solutions of the nonlinear Schrödinger equation $iu_t + \Delta u = \gamma |u|^\alpha u$. In particular, we develop a scattering theory in this class of solutions. As a consequence, we characterize self-similar solutions in $L^{\alpha+2}$ whose gradient is in L^2. In addition, we prove the existence of classical H^1 global solutions having various specified rates of decay as $t \to \infty$.

Key words. nonlinear Schrödinger equation, self-similar solution, scattering theory

AMS subject classifications. 35Q55, 35B40

PII. S0036141099351309

1. Introduction.

This paper is devoted to the study of a certain class of mild solutions of the nonlinear Schrödinger equation

$$iu_t + \Delta u = \gamma |u|^\alpha u, \tag{1.1}$$

where $u = u(t,x) \in \mathbb{C}$, $t > 0$, $x \in \mathbb{R}^N$, γ is a fixed real number, and $\alpha > 0$. As usual, a mild solution refers to a solution of the corresponding integral equation. Since we consider solutions of (1.1) for $t > 0$, i.e., without specifying an initial value, the corresponding integral equation has $u(\tau)$ as "initial value" for any $\tau > 0$. See Definition 3.1 for the precise formulation. In particular, the solution must be in $L^\infty_{\text{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$.

Historically, (1.1) has been studied in $H^1(\mathbb{R}^N)$, often with the additional restriction that α be "subcritical", i.e.,

$$\alpha < \frac{4}{N-2} \quad (\alpha < \infty \text{ if } N = 1,2).$$

In this case, the natural energy of a solution

$$E(u(t)) = \frac{1}{2}\int_{\mathbb{R}^N} |\nabla u(t,x)|^2\,dx + \frac{\gamma}{\alpha+2}\int_{\mathbb{R}^N} |u(t,x)|^{\alpha+2}\,dx$$

is well defined.

On the other hand, H^1 is not well adapted for the study of self-similar solutions of (1.1), i.e., solutions u of the form

$$u(t,x) = t^{-\frac{p}{2}} f\Big(\frac{x}{\sqrt{t}}\Big),$$

where $p \in \mathbb{C}$ and $\operatorname{Re} p = 2/\alpha$. Conservation laws for H^1 solutions of (1.1) combined with the dilation properties of self-similar solutions show that nontrivial H^1 self-similar solutions of (1.1) could only exist if $\alpha = 4/N$. And even if $\alpha = 4/N$, no radially

*Received by the editors February 5, 1999; accepted for publication June 24, 1999; published electronically February 28, 2000.

http://www.siam.org/journals/sima/31-3/35130.html

[†] Analyse Numérique—UMR CNRS 7598, Université Pierre et Marie Curie, 4 place Jussieu, F-75252 Paris Cedex 05, France (cazenave@ccr.jussieu.fr).

[‡] Laboratoire Analyse Géométrie et Applications, UMR CNRS 7539, Institut Galilée–Université Paris XIII, 99 Avenue J.-B. Clément, F-93430 Villetaneuse, France (weissler@math.univ-paris13.fr).

symmetric nontrivial H^1 self-similar solutions exist (see [7], [9]). As far as we are aware, however, the existence of nontrivial self-similar solutions with $\nabla f \in L^2(\mathbb{R}^N)$ and $f \in L^{\alpha+2}(\mathbb{R}^N)$ is an open question; see [10].

If $u(t,x) = t^{-\frac{p}{2}} f(\frac{x}{\sqrt{t}})$, where $\operatorname{Re} p = 2/\alpha$ and $f \in L^{\alpha+2}(\mathbb{R}^N)$, then

$$\|u(t)\|_{L^{\alpha+2}} = t^{-\beta}\|f\|_{L^{\alpha+2}},$$

where

$$\beta = \frac{4-(N-2)\alpha}{2\alpha(\alpha+2)}. \tag{1.2}$$

(Note that $\beta > 0$ precisely when α is subcritical.) Thus, we are motivated to consider the class of mild solutions u of (1.1) such that

$$u \in X_\alpha \equiv \{u \in L^\infty_{\text{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N));\ \sup_{t>0} t^\beta \|u(t)\|_{L^{\alpha+2}} < \infty\}. \tag{1.3}$$

We believe this to be a natural space in which to study solutions of (1.1). As we shall see, this study will yield new results about H^1 solutions.

In a previous article [4] we proved that if

$$\alpha_0 < \alpha < \frac{4}{N-2}, \tag{1.4}$$

where α_0 is the positive root of the polynomial $N\alpha^2 + (N-2)\alpha - 4$, and if $\varphi : \mathbb{R}^N \to \mathbb{C}$ is small enough in the space

$$W_\alpha = \{\varphi \in \mathcal{S}'(\mathbb{R}^N);\ \sup_{t>0} t^\beta \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} < \infty\}, \tag{1.5}$$

where

$$\mathcal{T}(t) = e^{it\Delta}$$

is the Schrödinger group on $\mathbb{R}^N$, then φ is the initial value of a (mild) solution $u \in X_\alpha$ of (1.1). If, in addition, $\varphi \in H^1(\mathbb{R}^N)$, then the resulting solution coincides with the "classical" H^1 solution. Furthermore, we showed that the resulting set of solutions includes a class of nontrivial self-similar solutions, and, if $\alpha < 4/N$, it includes a class of H^1 solutions which are asymptotically self-similar as $t \to \infty$.

In this paper, we continue the study of X_α solutions of (1.1). Section 2 gives the technical preliminaries, including information about X_α and W_α and related spaces. In section 3 below (Proposition 3.3) we prove that if α satisfies (1.4), then every solution $u \in X_\alpha$ of (1.1) has an initial value $\varphi \in W_\alpha$ and a scattering state $u^+ \in W_\alpha$. If, in addition, $\alpha > 4/N$ and u is an H^1 solution, then u^+ is also a scattering state in H^1 (Proposition 3.10). Furthermore, we obtain stronger estimates on the initial value, the scattering state, and decay of the integral term in case the solution $u \in X_\alpha$ in fact decays as $t \to \infty$ as $t^{-\mu}$ for some $\mu > \beta$. See Propositions 3.6, 3.7, and 3.8, Corollary 3.9, and Proposition 3.10 (iv).

In section 4 we show that the scattering state u^+ of the X_α solution u is (essentially) the Fourier transform of the initial value ψ of the function $v(s,y)$ related to $u(t,x)$ by the pseudoconformal transformation.

In section 5 we prove that if α satisfies (1.4) and u^+ is sufficiently small in W_α, then there exists a solution $u \in X_\alpha$ of (1.1) which has scattering state u^+. The results of this section, along with those of section 3 and [4], construct a "low energy" scattering theory for (1.1) in the space W_α. We also show (Proposition 5.3) that if, in addition, $\alpha > 4/N$ and $u^+ \in H^1(\mathbb{R}^N)$, then the resulting solution $u(t)$ is an H^1 solution. Moreover (Proposition 5.4), if α satisfies (1.4) and $u^+ \in H^1(\mathbb{R}^N) \cap W_\alpha$ is sufficiently small in W_α, and if $\|\mathcal{T}(t)u^+\|_{L^{\alpha+2}}$ decays like $t^{-\mu}$ for some $\mu > \beta$, then the resulting solution $u(t)$ is an H^1 solution. The smallness condition is not needed if $\gamma > 0$ or $\alpha < 4/N$ (Proposition 5.5).

We should point out that while the solution $u \in X_\alpha$ uniquely determines its initial value $\varphi \in W_\alpha$ and scattering state $u^+ \in W_\alpha$, the converse is not necessarily true. If u^+ (respectively, φ) is sufficiently small in W_α, then there is a solution $u \in X_\alpha$ with scattering state u^+ (respectively, with initial value φ). We know this solution to be unique among functions $u \in X_\alpha$ with $\sup_{t>0} t^\beta \|u(t)\|_{L^{\alpha+2}} \le M$ for some specific M determined by the parameters and W_α norm of u^+ (respectively, φ). It is conceivable (though we have no example) that another larger solution $\widetilde{u} \in X_\alpha$ of (1.1) will have the same scattering state (respectively, initial value). Indeed, for a solution $u \in X_\alpha$ of (1.1) we have not shown that $u^+ = 0$ (or $\varphi = 0$) implies $u \equiv 0$. See Proposition 3.8 for a partial result. Furthermore, we show in section 6 (Proposition 6.3) that if $u(t,x) = t^{-\frac{p}{2}} f(x/\sqrt{t})$ with $\operatorname{Re} p = 2/\alpha$ is a self-similar solution of (1.1), and if $f \in L^{\alpha+2}(\mathbb{R}^N)$ and $\nabla f \in L^2(\mathbb{R}^N)$, then its scattering state u^+ must be zero. In other words, if (1.4), a nontrivial finite energy self-similar solution of (1.1) would provide an example of a nontrivial X_α solution of (1.1) with $u^+ = 0$. Proposition 6.8 shows a partial converse: any radially symmetric self-similar solution $u \in X_\alpha$ of (1.1) (whose profile is also in a certain $L^q(\mathbb{R}^N)$; see the statement of Proposition 6.8) must be of finite energy.

The last section (before the appendix) is concerned with the comparison between scattering states and the asymptotic form of a solution as $t \to \infty$. As we shall see, different solutions of (1.1) in X_α can have the same asymptotic form as $t \to \infty$. On the other hand, at least among the solutions constructed by the "low energy" scattering theory in X_α, there is a one-to-one correspondence between solution and scattering state. Thus, the scattering state is a more precise description of the solution. In this section (section 7), we construct H^1 solutions of (1.1) which are asymptotic as $t \to \infty$ to various self-similar solutions of the linear Schrödinger equation. Indeed, if $\psi(x)$ is C^∞ away from the origin and homogeneous of degree $-q$, $0 < \operatorname{Re} q < N$, then the solution of the linear Schrödinger equation in $\mathcal{S}'(\mathbb{R}^N)$ given by $v(t) = \mathcal{T}(t)\psi$ is self-similar in that $\lambda^q v(\lambda^2 t, \lambda x) \equiv v(t,x)$. If we further require

$$\frac{\alpha+2}{\alpha+1} < \frac{N}{\operatorname{Re} q} < \alpha + 2,$$

then $\mathcal{T}(t)\psi \in L^{\alpha+2}(\mathbb{R}^N)$ for all $t > 0$, and so $v \in X_\alpha$. (See [4, 5, 12, 13]. This last reference shows that C^∞ can be replaced by lesser regularity.) Under the additional assumption that

$$\operatorname{Re} q > \min\Big\{\frac{N}{2}, \frac{2}{\alpha}\Big\},$$

we prove the existence of initial values and scattering states in $W_\alpha \cap H^1(\mathbb{R}^N)$ which give rise to solutions $u \in X_\alpha \cap C([0,\infty), H^1(\mathbb{R}^N))$, asymptotic as $t \to \infty$ to a linear self-similar solution of the above form. For such solutions, we are able to show that as $t \to \infty$

$$0 < a \le t^\nu \|u(t)\|_{L^{\alpha+2}} \le b < \infty$$

for certain $\nu > \beta$. We also obtain estimates on the difference $u(t) - \mathcal{T}(t)u^+$ in $L^{\alpha+2}(\mathbb{R}^N)$ and $H^1(\mathbb{R}^N)$ as $t \to \infty$. For similar results for the nonlinear heat equation, but in terms of the decay rate of $\|u(t)\|_{L^\infty}$, see Theorem 3.8 in [11].

We will have occasion to use the following property of the Schrödinger group on $\mathcal{S}'(\mathbb{R}^N)$: If $u_n \to u$ in $\mathcal{S}'(\mathbb{R}^N)$ and if $t_n \to t$, then $\mathcal{T}(t_n)u_n \to \mathcal{T}(t)u$ in $\mathcal{S}'(\mathbb{R}^N)$. Indeed, given $\theta \in \mathcal{S}(\mathbb{R}^N)$,

$$\langle \mathcal{T}(t_n)u_n, \theta\rangle_{\mathcal{S}',\mathcal{S}} = \langle u_n, \mathcal{T}(-t_n)\theta\rangle_{\mathcal{S}',\mathcal{S}} \to \langle u, \mathcal{T}(-t)\theta\rangle_{\mathcal{S}',\mathcal{S}} = \langle \mathcal{T}(t)u, \theta\rangle_{\mathcal{S}',\mathcal{S}},$$

since $u_n \to u$ in $\mathcal{S}'(\mathbb{R}^N)$ and $\mathcal{T}(-t_n)\theta \to \mathcal{T}(-t)\theta$ in $\mathcal{S}(\mathbb{R}^N)$ (see [14, Theorem 2.17, p. 52]).

2. Function spaces and preliminary estimates. In this section, we establish various estimates of the integral

$$\mathcal{G}(\tau, \sigma, t, u) = \int_\tau^\sigma \mathcal{T}(t-s)|u(s)|^\alpha u(s)\, ds. \tag{2.1}$$

Here, $\alpha > 0$, $0 \le \tau < \infty$, $0 \le \sigma \le \infty$, and $t \in \mathbb{R}$. All these estimates rely on the elementary property

$$\|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} \le |t|^{-\frac{N\alpha}{2(\alpha+2)}}\|\varphi\|_{L^{\frac{\alpha+2}{\alpha+1}}}, \tag{2.2}$$

which holds for all $t \ne 0$ and all $\varphi \in L^{\frac{\alpha+2}{\alpha+1}}(\mathbb{R}^N)$ and

$$L^{\frac{\alpha+2}{\alpha+1}}(\mathbb{R}^N) \hookrightarrow H^{-\frac{N\alpha}{2(\alpha+2)}}, \tag{2.3}$$

which follows by duality from the Sobolev embedding $H^{\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N) \hookrightarrow L^{\alpha+2}(\mathbb{R}^N)$. We begin with a simple observation.

Lemma 2.1. *The following properties hold:*

(i) *If $u \in L^{\alpha+1}_{\mathrm{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$ and if $0 < \tau \le \sigma < \infty$, then for every $t \in \mathbb{R}$ the integral in* (2.1) *is absolutely convergent in $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$ and*

$$\|\mathcal{G}(\tau,\sigma,t,u)\|_{H^{-\frac{N\alpha}{2(\alpha+2)}}} \le C\int_\tau^\sigma \|u(s)\|^{\alpha+1}_{L^{\alpha+2}}\, ds \tag{2.4}$$

for some constant $C = C(N,\alpha)$. If $u \in L^{\alpha+1}_{\mathrm{loc}}([0,\infty), L^{\alpha+2}(\mathbb{R}^N))$, then the integral is also convergent for $\tau = 0$ and the estimate (2.4) *holds.*

(ii) *Under the assumptions of* (i) *above, $\mathcal{G}(\tau,\sigma,t,u)$ depends continuously in $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$ on τ,σ,t. Moreover, for all $t,t' \in \mathbb{R}$, $\mathcal{T}(t')\mathcal{G}(\tau,\sigma,t,u) = \mathcal{G}(\tau,\sigma,t+t',u)$.*

(iii) *If $u \in L^{\alpha+1}_{\mathrm{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$ and if $0 < \tau \le \sigma < \infty$, then for every $t \notin [\tau,\sigma]$ the integral in* (2.1) *is absolutely convergent in $L^{\alpha+2}(\mathbb{R}^N)$. If $u \in L^{\alpha+1}_{\mathrm{loc}}([0,\infty), L^{\alpha+2}(\mathbb{R}^N))$, then the same property holds for $\tau = 0$. If we assume that $\alpha < 4/(N-2)$ (so that $\frac{N\alpha}{2(\alpha+2)} < 1$) and that $u \in L^\infty_{\mathrm{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$, then the integral is also convergent for every $t \in [\tau,\sigma]$ such that $t+\tau > 0$.*

(iv) *If u,v both satisfy the assumptions of* (iii) *above, then*

$$\begin{aligned} &\|\mathcal{G}(\tau,\sigma,t,u) - \mathcal{G}(\tau,\sigma,t,v)\|_{L^{\alpha+2}} \\ &\quad \le C\textstyle\int_\tau^\sigma |t-s|^{-\frac{N\alpha}{2(\alpha+2)}}(\|u(s)\|^\alpha_{L^{\alpha+2}} + \|v(s)\|^\alpha_{L^{\alpha+2}})\|u(s)-v(s)\|_{L^{\alpha+2}}\, ds, \end{aligned} \tag{2.5}$$

for some constant $C = C(\alpha)$.

Proof. By (2.3), $\| |u|^\alpha u\|_{H^{-\frac{N\alpha}{2(\alpha+2)}}} \le C\|u\|_{L^{\alpha+2}}^{\alpha+1}$. Properties (i) and (ii) follow immediately, since $(\mathcal{T}(t))_{t\in\mathbb{R}}$ is a group of isometries in $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$. The other properties are immediate consequences of (2.2) and Hölder's inequality. □

We next introduce the various spaces of initial values and time-dependent functions which we will use to study the Schrödinger equation (1.1). Suppose (1.2) and (1.4), so that

$$\beta(\alpha+1) < 1, \quad \frac{N\alpha}{2(\alpha+2)} < 1, \tag{2.6}$$

and

$$\frac{N\alpha}{2(\alpha+2)} + \alpha\beta = 1. \tag{2.7}$$

The spaces W_α and X_α defined by (1.5) and (1.3) with the norms

$$\|\varphi\|_{W_\alpha} = \sup_{t>0} t^\beta \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}}, \quad \|u\|_{X_\alpha} = \sup_{t>0} t^\beta \|u(t)\|_{L^{\alpha+2}}$$

are Banach spaces. We observe that $W_\alpha = \{\varphi \in \mathcal{S}'(\mathbb{R}^N);\ \mathcal{T}(\cdot)\varphi \in X_\alpha\}$. Moreover, it is clear that for all $t \ge 0$, $\mathcal{T}(t) : W_\alpha \to W_\alpha$ with

$$\|\mathcal{T}(t)\|_{\mathcal{L}(W_\alpha)} \le 1. \tag{2.8}$$

Given α as above and $\mu \ge 0$, we also define

$$\mathcal{W}_{\alpha,\mu} = \{\varphi \in \mathcal{S}'(\mathbb{R}^N);\ \sup_{t>0} t^\mu \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} < \infty\} \tag{2.9}$$

and

$$\mathcal{X}_{\alpha,\mu} = \{u \in L^\infty_{\mathrm{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N));\ \sup_{t>0} t^\mu \|u(t)\|_{L^{\alpha+2}} < \infty\}. \tag{2.10}$$

It is clear that $\mathcal{W}_{\alpha,\mu}$ and $\mathcal{X}_{\alpha,\mu}$ are Banach spaces with the norms

$$\|\varphi\|_{\mathcal{W}_{\alpha,\mu}} = \sup_{t>0} t^\mu \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} \quad \text{and} \quad \|u\|_{\mathcal{X}_{\alpha,\mu}} = \sup_{t>0} t^\mu \|u(t)\|_{L^{\alpha+2}}.$$

We observe that $\mathcal{W}_{\alpha,\mu} = \{\varphi \in \mathcal{S}'(\mathbb{R}^N);\ \mathcal{T}(\cdot)\varphi \in \mathcal{X}_{\alpha,\mu}\}$. Moreover, it is clear that for all $t \ge 0$, $\mathcal{T}(t) : \mathcal{W}_{\alpha,\mu} \to \mathcal{W}_{\alpha,\mu}$ with

$$\|\mathcal{T}(t)\|_{\mathcal{L}(\mathcal{W}_{\alpha,\mu})} \le 1. \tag{2.11}$$

It is also clear that $\mathcal{W}_{\alpha,\beta} = W_\alpha$ and $\mathcal{X}_{\alpha,\beta} = X_\alpha$, where β is defined by (2.6).

Remark 2.2. Note that the definitions of W_α and $\mathcal{W}_{\alpha,\mu}$ make sense for any $\alpha > 0$. On the other hand, if $\alpha > 4/(N-2)$, then $\beta < 0$; thus $W_\alpha = \{0\}$ because $\|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} \to 0$ as $t \downarrow 0$. Moreover, if $\alpha < \alpha_0$ then Proposition 8.1 also implies that $W_\alpha = \{0\}$. This result has been proved independently by Oru [12, Theorem V.2–5]. Furthermore, if $\mu > \frac{N\alpha}{2(\alpha+2)}$, then Proposition 8.1 implies that $\mathcal{W}_{\alpha,\mu} = \{0\}$.

We next estimate $\mathcal{G}(\tau,\sigma,t,u)$ for $u \in X_\alpha$ in the case $\sigma < \infty$.

Proposition 2.3. *If* (1.2) *and* (1.4), *then the following properties hold:*

(i) *For all* $u \in X_\alpha$ *and all* $0 \le \tau \le \sigma < \infty$ *and* $t \in \mathbb{R}$*, the integral in* (2.1) *is absolutely convergent in* $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$*. If* $|t| + \tau > 0$*, it is also absolutely convergent in* $L^{\alpha+2}(\mathbb{R}^N)$*. Moreover,*

$$\mathcal{T}(t')\mathcal{G}(\tau, \sigma, t, u) = \mathcal{G}(\tau, \sigma, t' + t, u) \tag{2.12}$$

for all $0 \le \tau \le \sigma < \infty$ *and all* $t, t' \in \mathbb{R}$*. Furthermore,* $\mathcal{G}(\tau, \sigma, t, u)$ *depends continuously in* $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$ *on* τ, σ, t *as above.*

(ii) *If* $u, v \in X_\alpha$ *and if* $u - v \in \mathcal{X}_{\alpha,\mu}$ *for some* $\mu \in (0, 1 - \alpha\beta)$*, then* $\mathcal{G}(\tau, \sigma, t, u) - \mathcal{G}(\tau, \sigma, t, v) \in \mathcal{W}_{\alpha,\mu}$ *for all* $\tau, \sigma, t \ge 0$*. Moreover,* $\mathcal{G}(\tau, \sigma, \cdot, u) - \mathcal{G}(\tau, \sigma, \cdot, v) \in \mathcal{X}_{\alpha,\mu}$ *and*

$$\begin{aligned} &\|\mathcal{G}(\tau, \sigma, t, u) - \mathcal{G}(\tau, \sigma, t, v)\|_{\mathcal{W}_{\alpha,\mu}} + \|\mathcal{G}(\tau, \sigma, \cdot, u) - \mathcal{G}(\tau, \sigma, \cdot, v)\|_{\mathcal{X}_{\alpha,\mu}} \\ &\qquad \le C(\|u\|^\alpha_{X_\alpha} + \|v\|^\alpha_{X_\alpha})\|u - v\|_{\mathcal{X}_{\alpha,\mu}}, \end{aligned} \tag{2.13}$$

where $C = C(N, \alpha, \mu)$*.*

Proof. By (2.6), property (i) follows from Lemma 2.1. Now let $\mu \in (0, 1 - \alpha\beta)$ and let $u, v \in X_\alpha$ be such that $u - v \in \mathcal{X}_{\alpha,\mu}$. Given $\theta \ge 0$, it follows from (2.12) that $\mathcal{T}(\theta)\mathcal{G}(\tau, \sigma, t, u) = \mathcal{G}(\tau, \sigma, \theta + t, u)$. Therefore, applying (2.5), we obtain

$$\begin{aligned} &\|\mathcal{T}(\theta)(\mathcal{G}(\tau, \sigma, t, u) - \mathcal{G}(\tau, \sigma, t, v))\|_{L^{\alpha+2}} \\ &\qquad \le C(\|u\|^\alpha_{X_\alpha} + \|v\|^\alpha_{X_\alpha})\|u - v\|_{\mathcal{X}_{\alpha,\mu}} \Big| \int_\tau^\sigma |\theta + t - s|^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\alpha\beta-\mu}\, ds \Big|. \end{aligned} \tag{2.14}$$

On the other hand, we deduce from (2.7) and (2.6) that

$$\begin{aligned} \int_0^\infty |\theta + t - s|^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\alpha\beta-\mu}\, ds &= (\theta + t)^{-\mu} \int_0^\infty |1 - s|^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\alpha\beta-\mu}\, ds \\ &= (\theta + t)^{-\mu} C(N, \alpha, \mu). \end{aligned} \tag{2.15}$$

Property (ii) follows from (2.14) and (2.15). □

We now estimate $\mathcal{G}(\tau, \infty, t, u)$ for $u \in X_\alpha$.

Proposition 2.4. *If* (1.2) *and* (1.4)*, then the following properties hold:*

(i) *Given* $u \in X_\alpha$*, the integral in* (2.1) *with* $\sigma = \infty$ *is absolutely convergent in* $L^{\alpha+2}(\mathbb{R}^N)$ *for all* $0 \le \tau < \infty$ *and* $t \in \mathbb{R}$ *such that* $|t| + \tau > 0$*. For* $t = \tau = 0$*, it converges in* $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N) + L^{\alpha+2}(\mathbb{R}^N)$*. Moreover,*

$$\mathcal{G}(\tau, \infty, t, u) - \mathcal{G}(\tau, \sigma, t, u) = \mathcal{G}(\sigma, \infty, t, u) \underset{\sigma \to \infty}{\longrightarrow} 0 \tag{2.16}$$

in $L^{\alpha+2}(\mathbb{R}^N)$ *for* $\tau \ge 0$ *and* $t \in \mathbb{R}$*, and*

$$\mathcal{T}(t')\mathcal{G}(\tau, \infty, t, u) = \mathcal{G}(\tau, \infty, t' + t, u) \tag{2.17}$$

for all $0 \le \tau < \infty$ *and all* $t, t' \in \mathbb{R}$*.*

(ii) *If* $u, v \in X_\alpha$ *and if* $u - v \in \mathcal{X}_{\alpha,\mu}$ *for some* $\mu \in (0, 1 - \alpha\beta)$*, then* $\mathcal{G}(\tau, \infty, t, u) - \mathcal{G}(\tau, \infty, t, v) \in \mathcal{W}_{\alpha,\mu}$ *for all* $\tau, t \ge 0$*. Moreover,* $\mathcal{G}(\tau, \infty, \cdot, u) - \mathcal{G}(\tau, \infty, \cdot, v) \in \mathcal{X}_{\alpha,\mu}$ *and*

$$\begin{aligned} &\|\mathcal{G}(\tau, \infty, t, u) - \mathcal{G}(\tau, \infty, t, v)\|_{\mathcal{W}_{\alpha,\mu}} + \|\mathcal{G}(\tau, \infty, \cdot, u) - \mathcal{G}(\tau, \infty, \cdot, v)\|_{\mathcal{X}_{\alpha,\mu}} \\ &\qquad \le C(\|u\|^\alpha_{X_\alpha} + \|v\|^\alpha_{X_\alpha})\|u - v\|_{\mathcal{X}_{\alpha,\mu}}, \end{aligned} \tag{2.18}$$

where $C = C(N, \alpha, \mu)$*.*

(iii) *If* $u, v \in X_\alpha$ *and if* $u - v \in \mathcal{X}_{\alpha,\mu}$ *for some* $\mu > 0$*, then*

$$\|\mathcal{G}(\,\cdot\,, \infty, \,\cdot\,, u) - \mathcal{G}(\,\cdot\,, \infty, \,\cdot\,, v)\|_{\mathcal{X}_{\alpha,\mu}} \le C(\|u\|^\alpha_{X_\alpha} + \|v\|^\alpha_{X_\alpha})\|u - v\|_{\mathcal{X}_{\alpha,\mu}}, \tag{2.19}$$

where $C = C(N, \alpha, \mu)$.

Proof. Let $u \in X_\alpha$. We first observe that

$$\begin{aligned}(2.20)\qquad &\|\mathcal{T}(t-s)|u(s)|^\alpha u(s)\|_{L^{\alpha+2}} \\ &\quad \le |t-s|^{-\frac{N\alpha}{2(\alpha+2)}}\|u(s)\|_{L^{\alpha+2}}^{\alpha+1} \le |t-s|^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\alpha\beta-\beta}\|u\|_{X_\alpha}^{\alpha+1}.\end{aligned}$$

It follows from (2.7) that $\|\mathcal{T}(t-s)|u(s)|^\alpha u(s)\|_{L^{\alpha+2}}$ is integrable as $s \to \infty$. The other singularities are $|t-s|^{-\frac{N\alpha}{2(\alpha+2)}}$ at $s = t$ and $s^{-\alpha\beta-\beta}$ at $s = 0$. By (2.6) both are integrable provided $t \neq 0$. Thus we see that the integral in (2.1) is absolutely convergent in $L^{\alpha+2}(\mathbb{R}^N)$ for all $0 \le \tau < \infty$ and $t \in \mathbb{R}$ such that $|t| + \tau > 0$. For $t = \tau = 0$, we write

$$\int_0^\infty \mathcal{T}(-s)|u(s)|^\alpha u(s)\,ds = \int_0^1 \mathcal{T}(-s)|u(s)|^\alpha u(s)\,ds + \int_1^\infty \mathcal{T}(-s)|u(s)|^\alpha u(s)\,ds.$$

The second integral on the right-hand side is absolutely convergent in $L^{\alpha+2}(\mathbb{R}^N)$ by the previous argument, and the first one is absolutely convergent in $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$ by Proposition 2.3 (i). This proves the first part of (i), and (2.16) is an immediate consequence. To prove (2.17), we observe that on the one hand,

$$\mathcal{T}(t')\mathcal{G}(\tau,\sigma,t,u) \underset{\sigma\to\infty}{\longrightarrow} \mathcal{T}(t')\mathcal{G}(\tau,\infty,t,u)$$

in $\mathcal{S}'(\mathbb{R}^N)$ by (2.16). On the other hand, by (2.12) and (2.16),

$$\mathcal{T}(t')\mathcal{G}(\tau,\sigma,t,u) = \mathcal{G}(\tau,\sigma,t'+t,u) \underset{\sigma\to\infty}{\longrightarrow} \mathcal{G}(\tau,\infty,t'+t,u)$$

in $\mathcal{S}'(\mathbb{R}^N)$.

Turning now to (ii), we let $u, v \in X_\alpha$ be such that $u - v \in \mathcal{X}_{\alpha,\mu}$ for some $\mu \in (0, 1-\alpha\beta)$. By (2.17) and (2.16), we may let $\sigma \to \infty$ in (2.13) and we obtain (2.18).

Finally, let $u, v \in X_\alpha$ be such that $u - v \in \mathcal{X}_{\alpha,\mu}$ for some $\mu > 0$. It follows from (2.2) and Hölder's inequality that

$$\begin{aligned}&\|\mathcal{G}(t,\infty,t,u) - \mathcal{G}(t,\infty,t,v)\|_{L^{\alpha+2}} \\ &\qquad \le C(\|u\|_{X_\alpha}^\alpha + \|v\|_{X_\alpha}^\alpha)\|u-v\|_{\mathcal{X}_{\alpha,\mu}} \int_t^\infty |t-s|^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\alpha\beta-\mu}\,ds \\ &\qquad = Ct^{-\mu}(\|u\|_{X_\alpha}^\alpha + \|v\|_{X_\alpha}^\alpha)\|u-v\|_{\mathcal{X}_{\alpha,\mu}} \int_1^\infty |1-s|^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\alpha\beta-\mu}\,ds.\end{aligned}$$

This proves property (iii). □

3. Mild solutions of the nonlinear Schrödinger equation in $L^{\alpha+2}(\mathbb{R}^N)$.

Throughout this paper, we consider solutions of (1.1) in the following sense.

Definition 3.1. *Given* $u \in L^\infty_{\mathrm{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$, *we say that* u *is a solution of* (1.1) *if*

$$(3.1)\qquad u(t) = \mathcal{T}(t-\tau)u(\tau) - i\gamma\int_\tau^t \mathcal{T}(t-s)|u(s)|^\alpha u(s)\,ds$$

for all $t, \tau > 0$.

Remark 3.2. At first sight, it seems that Definition 3.1 does not make sense since u is only defined almost everywhere; and so (3.1) could only be required to hold for almost all $t, \tau > 0$. One can show, however, that if $u \in L^\infty_{\text{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$ satisfies (3.1) for almost all $t, \tau > 0$, then u can be modified on a set of measure 0 in t so that $u \in C((0,\infty), \mathcal{S}'(\mathbb{R}^N))$ and satisfies (3.1) for all $t, \tau > 0$. Indeed, given any $\tau > 0$, it follows from Lemma 2.1 that the right-hand side of (3.1) is a continuous function of $t > 0$ with values in $\mathcal{S}'(\mathbb{R}^N)$. Fix $\tau_0 > 0$ such that (3.1) holds for almost all $t > 0$ and redefine u to be equal everywhere to the right-hand side of (3.1) with $\tau = \tau_0$. Given now any $t, \tau > 0$, using (3.1) with (τ, t) replaced successively by (τ_0, τ) and (τ_0, t), we see that u satisfies (3.1) at (t, τ). It is understood that we always consider this continuous representative of the solution u.

Proposition 3.3. *Suppose* (1.2) *and* (1.4). *If $u \in X_\alpha$ is a solution of* (1.1), *then the following properties hold:*

(i) *$u(t) \in W_\alpha$ for all $t > 0$ and $\sup_{t>0} \|u(t)\|_{W_\alpha} \le \|u\|_{X_\alpha} + C\|u\|_{X_\alpha}^{\alpha+1}$.*

(ii) *$\mathcal{T}(-t)u(t) \in W_\alpha$ for all $t > 0$ and $\sup_{t>0} \|\mathcal{T}(-t)u(t)\|_{W_\alpha} \le \|u\|_{X_\alpha} + C\|u\|_{X_\alpha}^{\alpha+1}$.*

(iii) *There exists $\varphi \in W_\alpha$ such that $u(t) \to \varphi$ as $t \downarrow 0$ in $\mathcal{S}'(\mathbb{R}^N)$ and*

$$u(t) = \mathcal{T}(t)\varphi - i\gamma \int_0^t \mathcal{T}(t-s)|u(s)|^\alpha u(s)\, ds \tag{3.2}$$

for all $t \ge 0$. The integral in (3.2) *makes sense in $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$. Furthermore,*

$$\|u(t) - \mathcal{T}(t)\varphi\|_{H^{-\frac{N\alpha}{2(\alpha+2)}}} \le Ct^{1-\beta(\alpha+1)}\|u\|_{X_\alpha}^{\alpha+1} \tag{3.3}$$

for all $t \ge 0$. Moreover, $u : [0,\infty) \to \mathcal{S}'(\mathbb{R}^N)$ is continuous and $u : (0,\infty) \to L^{\alpha+2}(\mathbb{R}^N)$ is weakly continuous.

(iv) *There exists $u^+ \in W_\alpha$ such that $\mathcal{T}(-t)u(t) \to u^+$ as $t \to \infty$ in $\mathcal{S}'(\mathbb{R}^N)$ and equation*

$$\mathcal{T}(\tau - t)u(t) = \mathcal{T}(\tau)u^+ + i\gamma \int_t^\infty \mathcal{T}(\tau - s)|u(s)|^\alpha u(s)\, ds \tag{3.4}$$

holds for all $t, \tau \ge 0$. If $t + \tau > 0$, the integral in (3.4) *makes sense in $L^{\alpha+2}(\mathbb{R}^N)$. (If $t = \tau = 0$, the integral makes sense in $L^{\alpha+2}(\mathbb{R}^N) + H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$.) Finally,*

$$\|\mathcal{T}(-t)u(t) - u^+\|_{L^{\alpha+2}} \le Ct^{-\beta}\|u\|_{X_\alpha}^{\alpha+1} \tag{3.5}$$

for all $t > 0$.

Proof. We begin by showing statement (ii). It follows from (3.1) that

$$\mathcal{T}(t-\tau)u(\tau) = u(t) + i\gamma \int_\tau^t \mathcal{T}(t-s)|u(s)|^\alpha u(s)\, ds. \tag{3.6}$$

Consider a fixed $\tau > 0$. It follows from (3.6) and (2.13) applied with $v = 0$, $\sigma = t$, and $\mu = \beta$ that

$$\|\mathcal{T}(\cdot - \tau)u(\tau)\|_{X_\alpha} \le \|u\|_{X_\alpha} + \gamma\|\mathcal{G}(\tau, \cdot, \cdot, u)\|_{X_\alpha} \le \|u\|_{X_\alpha}(1 + C\|u\|_{X_\alpha}^\alpha).$$

Since $\|\mathcal{T}(\cdot - \tau)u(\tau)\|_{X_\alpha} = \|\mathcal{T}(-\tau)u(\tau)\|_{W_\alpha}$, we deduce statement (ii). Property (i) follows from property (ii) and (2.8).

We turn to property (iii). We deduce from (3.1) and (2.12) that

$$\mathcal{T}(-t)u(t) - \mathcal{T}(-\tau)u(\tau) = -i\gamma \int_\tau^t \mathcal{T}(-s)|u(s)|^\alpha u(s)\, ds. \tag{3.7}$$

Next, we let $\tau \downarrow 0$. The right-hand side converges to an element of $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$ by Lemma 2.1. Since $\mathcal{T}(-t)u(t)$ is a fixed element of $\mathcal{S}'(\mathbb{R}^N)$, we see that there exists $\varphi \in \mathcal{S}'(\mathbb{R}^N)$ such that $\mathcal{T}(-\tau)u(\tau) \to \varphi$ in $\varphi \in \mathcal{S}'(\mathbb{R}^N)$ as $\tau \downarrow 0$. This implies that $u(\tau) \to \varphi$ in $\varphi \in \mathcal{S}'(\mathbb{R}^N)$ as $\tau \downarrow 0$. We then find

$$\mathcal{T}(-t)u(t) = \varphi - i\gamma \int_0^t \mathcal{T}(-s)|u(s)|^\alpha u(s)\, ds \tag{3.8}$$

for all $t \geq 0$. We may then apply $\mathcal{T}(t)$ and we obtain (3.2). Here again, the right-hand side makes sense in $H^{-\frac{N\alpha}{2(\alpha+2)}}(\mathbb{R}^N)$. Furthermore, it follows from (2.4) that

$$\|u(t) - \mathcal{T}(t)\varphi\|_{H^{-\frac{N\alpha}{2(\alpha+2)}}} \leq C\|u\|_{X_\alpha}^{\alpha+1} \int_0^t s^{-(\alpha+1)\beta}\, ds,$$

and the estimate (3.3) follows. Using (3.2) and the estimate (2.13) with $\tau = 0$, $\sigma = t$, $\mu = \beta$, and $v = 0$, we see that $\varphi \in W_\alpha$ and $\|\varphi\|_{W_\alpha} \leq \|u\|_{W_\alpha} + C\|u\|_{W_\alpha}^{\alpha+1}$. Since $\mathcal{T}(\cdot)\varphi : [0,\infty) \to \mathcal{S}'(\mathbb{R}^N)$ is continuous, it follows from Proposition 2.3 (i) that $u : [0,\infty) \to \mathcal{S}'(\mathbb{R}^N)$ is continuous, and this proves (iii).

We now construct u^+. We fix $\tau > 0$ and we let $t \to \infty$ in (3.7). It follows from (2.16) that the right-hand side of (3.7) converges to $-i\gamma\mathcal{G}(\tau, \infty, 0, u)$ in $L^{\alpha+2}(\mathbb{R}^N)$ as $t \to \infty$, and thus $\mathcal{T}(-t)u(t)$ converges in $\mathcal{S}'(\mathbb{R}^N)$ to a limit that we call u^+. Thus we obtain (3.4) with $\tau = 0$. The general case of (3.4) now follows from applying $\mathcal{T}(\tau)$ (which is possible by (2.17)). It follows from (3.4) applied with $\tau = 0$ and Proposition 2.4 (ii) that $u^+ \in W_\alpha$. Finally, since

$$\begin{aligned} &\Big\|\int_t^\infty \mathcal{T}(-s)|u(s)|^\alpha u(s)\, ds\Big\|_{L^{\alpha+2}} \\ &\qquad \leq C\|u\|_{X_\alpha}^{\alpha+1} \int_t^\infty s^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\beta(\alpha+1)}\, ds \leq C\|u\|_{X_\alpha}^{\alpha+1} t^{-\beta}, \end{aligned} \tag{3.9}$$

we obtain (3.5). This proves (iv). □

REMARK 3.4. It follows from (2.13) and (2.18) that φ and u^+ depend continuously on u. More precisely, given two solutions u_1 and u_2 of (1.1),

$$\|\varphi_1 - \varphi_2\|_{W_\alpha} + \|u_1^+ - u_2^+\|_{W_\alpha} \leq \|u_1 - u_2\|_{X_\alpha} + C(\|u_1\|_{X_\alpha}^\alpha + \|u_2\|_{X_\alpha}^\alpha)\|u_1 - u_2\|_{X_\alpha}, \tag{3.10}$$

where φ_j and u_j^+ are the corresponding initial values and scattering states.

REMARK 3.5. In principle, we could have considered solutions of (1.1) in X_α without requiring $\alpha > \alpha_0$. It is not too hard to check that Proposition 3.3 would still be true, except for the construction of the initial value φ. The reason for this is that the condition $\beta(\alpha + 1) < 1$ is needed only for the convergence of integrals at 0. In particular, $u^+ \in W_\alpha$ and $u(t) \in W_\alpha$ for all $t > 0$. On the other hand, if $\alpha < \alpha_0$, then $W_\alpha = \{0\}$ by Remark 2.2. Thus, if $\alpha < \alpha_0$ then (1.1) has no nontrivial solutions in X_α and in particular no nontrivial self-similar solutions with profile in $L^{\alpha+2}(\mathbb{R}^N)$.

PROPOSITION 3.6. *Suppose* (1.2) *and* (1.4) *and let* $\mu > 0$. *If* $u_1, u_2 \in X_\alpha$ *are two solutions of* (1.1) *with initial values* $\varphi_1, \varphi_2 \in W_\alpha$ *and scattering states* $u_1^+, u_2^+ \in W_\alpha$, *respectively, and if in addition* $u_1 - u_2 \in \mathcal{X}_{\alpha,\mu}$, *then* $u_1^+ - u_2^+ \in \mathcal{W}_{\alpha,\mu}$ *and*

$$\|u_1^+ - u_2^+\|_{\mathcal{W}_{\alpha,\mu}} \leq \|u_1 - u_2\|_{\mathcal{X}_{\alpha,\mu}} + C(\|u_1\|_{X_\alpha}^\alpha + \|u_2\|_{X_\alpha}^\alpha)\|u_1 - u_2\|_{\mathcal{X}_{\alpha,\mu}} \tag{3.11}$$

for some constant C *independent of* u_1 *and* u_2. *If, furthermore,* $\alpha\beta + \mu < 1$, *then* $\varphi_1 - \varphi_2 \in \mathcal{W}_{\alpha,\mu}$ *and* $u_1(t) - u_2(t) \in \mathcal{W}_{\alpha,\mu}$ *for all* $t > 0$, *and*

$$\begin{aligned} &\max\{\|\varphi_1 - \varphi_2\|_{\mathcal{W}_{\alpha,\mu}}, \sup_{t>0} \|u_1(t) - u_2(t)\|_{\mathcal{W}_{\alpha,\mu}}\} \\ &\qquad \leq \|u_1 - u_2\|_{\mathcal{X}_{\alpha,\mu}} + C(\|u_1\|_{X_\alpha}^\alpha + \|u_2\|_{X_\alpha}^\alpha)\|u_1 - u_2\|_{\mathcal{X}_{\alpha,\mu}} \end{aligned} \tag{3.12}$$

for some constant C independent of u_1 and u_2.

Proof. It follows from (3.4) with $\tau = t$ that

$$\mathcal{T}(t)(u_1^+ - u_2^+) = u_1(t) - u_2(t) - i\gamma(\mathcal{G}(t,\infty,t,u_1) - \mathcal{G}(t,\infty,t,u_2)),$$

so (3.11) follows from (2.19). Next, assume $\alpha\beta + \mu < 1$ and consider (3.4) with a fixed value of $t > 0$, i.e.,

$$\mathcal{T}(\cdot)\mathcal{T}(-t)(u_1(t) - u_2(t)) = \mathcal{T}(\cdot)(u_1^+ - u_2^+) + i\gamma(\mathcal{G}(t,\infty,\cdot,u_1) - \mathcal{G}(t,\infty,\cdot,u_2)).$$

It follows from (3.11) and (2.18) that

$$\|\mathcal{T}(-t)(u_1(t) - u_2(t))\|_{\mathcal{W}_{\alpha,\mu}} \le \|u_1 - u_2\|_{\mathcal{X}_{\alpha,\mu}} + C(\|u_1\|_{X_\alpha}^\alpha + \|u_2\|_{X_\alpha}^\alpha)\|u_1 - u_2\|_{\mathcal{X}_{\alpha,\mu}}.$$

We deduce the second estimate in (3.12) by applying (2.11). Next, it follows from (3.2) that $\mathcal{T}(t)(\varphi_1 - \varphi_2) = u_1(t) - u_2(t) + i\gamma(\mathcal{G}(0,t,t,u_1) - \mathcal{G}(0,t,t,u_2))$; and so the first estimate in (3.12) is a consequence of (2.13). □

PROPOSITION 3.7. *Suppose* (1.2) *and* (1.4). *Let $u \in X_\alpha$ be a solution of* (1.1) *with initial value $\varphi \in W_\alpha$ and scattering state $u^+ \in W_\alpha$. If $u \in \mathcal{X}_{\alpha,\mu}$ for some $\mu > \beta$, then the following properties hold:*

(i) $t^\mu \|u(t) - \mathcal{T}(t)u^+\|_{L^{\alpha+2}} \le C\|u\|_{\mathcal{X}_{\alpha,\mu}}^{\alpha+1} t^{-\alpha(\mu-\beta)}$ *for all $t > 0$.*

(ii) $t^\mu \|u(t) - \mathcal{T}(t)\varphi\|_{L^{\alpha+2}} \le C\|u\|_{\mathcal{X}_{\alpha,\mu}}^{\alpha+1} t^{-\alpha(\mu-\beta)}$ *for all $t > 0$, provided $(\alpha+1)\mu < 1$.*

(iii) $t^\mu \|\mathcal{T}(-t)u(t) - u^+\|_{L^{\alpha+2}} \le C\|u\|_{\mathcal{X}_{\alpha,\mu}}^{\alpha+1} t^{-\alpha(\mu-\beta)}$ *for all $t > 0$.*

(iv) $\|\mathcal{T}(-t)u(t) - u^+\|_{H^{-\frac{N\alpha}{2(\alpha+2)}}} \le C\|u\|_{\mathcal{X}_{\alpha,\mu}}^{\alpha+1} t^{1-(\alpha+1)\mu}$ *for all $t > 0$, provided $(\alpha + 1)\mu > 1$.*

(v) $t^\mu \|\mathcal{T}(t)(\varphi - u^+)\|_{L^{\alpha+2}} \le C\|u\|_{\mathcal{X}_{\alpha,\mu}}^{\alpha+1} t^{-\alpha(\mu-\beta)}$ *for all $t > 0$, provided $(\alpha + 1)\mu < 1$.*

Proof. All these results follow from the various integral equations, where in the integral term $\|u(s)\|_{L^{\alpha+2}}$ is always estimated by $s^{-\mu}\|u\|_{\mathcal{X}_{\alpha,\mu}}$. □

PROPOSITION 3.8. *Suppose* (1.2) *and* (1.4). *Let $u \in X_\alpha$ be a solution of* (1.1) *with scattering state $u^+ \in W_\alpha$. If $u \in \mathcal{X}_{\alpha,\mu}$ for some $\mu > \beta$ and if $u^+ = 0$, then $u(t) \equiv 0$.*

Proof. We apply (3.4) with $\tau = t$ and we deduce that

$$\|u(t)\|_{L^{\alpha+2}} \le C \int_t^\infty (s-t)^{-\frac{N\alpha}{2(\alpha+2)}} \|u(s)\|_{L^{\alpha+2}}^{\alpha+1}\, ds;$$

and so, setting $f(t) = \sup_{s \ge t} \|u(s)\|_{L^{\alpha+2}}$,

$$\|u(t)\|_{L^{\alpha+2}} \le Cf(t) \int_t^\infty (s-t)^{-\frac{N\alpha}{2(\alpha+2)}} s^{-\mu\alpha}\, ds = Cf(t)t^{-\alpha(\mu-\beta)} \tag{3.13}$$

by (2.7). Since the right-hand side of (3.13) is nonincreasing, we see that $f(t) \le Cf(t)t^{-\alpha(\mu-\beta)}$. Thus $f(t) = 0$ for t sufficiently large. Fix now t_0 large enough so that $u(t) = 0$ for $t \ge t_0$. We deduce from (3.6) that

$$u(t) = i\gamma \int_t^{t_0} \mathcal{T}(t-s)|u(s)|^\alpha u(s)\, ds \tag{3.14}$$

for all $t \ge 0$. Fix $0 < \varepsilon < t_0$. Since $\sup_{\varepsilon \le s \le t_0} \|u(s)\|_{L^{\alpha+2}} < \infty$, we deduce from (3.14) that

$$\|u(t)\|_{L^{\alpha+2}} \le C \int_t^{t_0} |t-s|^{-\frac{N\alpha}{2(\alpha+2)}} \|u(s)\|_{L^{\alpha+2}}\, ds$$

for $\varepsilon \le t \le t_0$, and by a singular Gronwall-type argument, $u(t) = 0$ for all $\varepsilon \le t \le t_0$. The result follows by letting $\varepsilon \downarrow 0$. □

COROLLARY 3.9. *Suppose* (1.2) *and* (1.4). *If* $\mu > \frac{N\alpha}{2(\alpha+2)}$ *and if* u *is a solution of* (1.1) *in* $X_\alpha \cap \mathcal{X}_{\alpha,\mu}$, *then* $u \equiv 0$.

Proof. By Proposition 3.6, $u^+ \in \mathcal{W}_{\alpha,\mu}$. Since $\mathcal{W}_{\alpha,\mu} = \{0\}$ by Remark 2.2, the result follows from Proposition 3.8. □

We conclude this section by considering solutions of (1.1) that are both in X_α and $H^1(\mathbb{R}^N)$.

PROPOSITION 3.10. *Suppose* (1.2) *and* (1.4) *and let* $u \in X_\alpha$ *be a solution of* (1.1) *in the sense of Definition* 3.1 *with scattering state* u^+. *If* $u(t_0) \in H^1(\mathbb{R}^N)$ *for some* $t_0 > 0$, *then the following properties hold:*

(i) $u(t) \in H^1(\mathbb{R}^N)$ *for all* $t > 0$ *and* u *is the "classical"* H^1 *solution determined by* $u(t_0)$.

(ii) $u^+ \in H^1(\mathbb{R}^N)$.

(iii) *If* $\alpha > 4/N$, *or if* $\alpha = 4/N$ *and* $\|u(t_0)\|_{H^1}$ *is sufficiently small, then* $\mathcal{T}(-t)u(t) \to u^+$ *in* $H^1(\mathbb{R}^N)$ *as* $t \to \infty$ *and* $\|u(t) - \mathcal{T}(t)u^+\|_{H^1} \le Ct^{-\beta(\alpha+2)\frac{N\alpha-4}{N\alpha}}$.

(iv) *If* $u \in \mathcal{X}_{\alpha,\mu}$ *for some* $\mu > \beta$, *then* $\mathcal{T}(-t)u(t) \to u^+$ *in* $H^1(\mathbb{R}^N)$ *as* $t \to \infty$ *and* $\|u(t) - \mathcal{T}(t)u^+\|_{H^1} \le Ct^{-\alpha(\mu-\beta)}$.

Proof. Let v be the H^1 solution of (1.1) with the initial value $v(t_0) = u(t_0)$, which is defined on the maximal interval $(t_0 - T_*, t_0 + T^*)$. It follows that

$$u(t) - v(t) = -i\mathcal{G}(t_0, t, t, u) + i\mathcal{G}(t_0, t, t, v)$$

for $t_0 < t < t_0 + T^*$. By applying (2.5), we deduce that

$$\begin{aligned} &\|u(t) - v(t)\|_{L^{\alpha+2}} \\ &\quad \le C \sup_{t_0 \le s \le t} (\|u(s)\|_{L^{\alpha+2}} + \|v(s)\|_{L^{\alpha+2}})^\alpha \int_{t_0}^t |t-s|^{-\frac{N\alpha}{2(\alpha+2)}} \|u(s) - v(s)\|_{L^{\alpha+2}}\, ds. \end{aligned}$$

It now follows from a singular Gronwall argument that $u(t) = v(t)$ for $t_0 \le t < t_0 + T^*$. Since $\sup_{t \ge t_0} \|u(t)\|_{L^{\alpha+2}} < \infty$, it follows from conservation of energy that $\sup_{t_0 \le t < t_0 + T^*} \|u(t)\|_{H^1} < \infty$, so that $T^* = \infty$. For $t_0 - T_* < t < t_0$, we have $u(t) - v(t) = i\mathcal{G}(t, t_0, t, u) + i\mathcal{G}(t, t_0, t, v)$ and a similar argument shows that $T_* \ge t_0$ and that $v(t) = u(t)$ for all $0 < t < t_0$. This proves (i).

We now observe that $\|u(t)\|_{L^{\alpha+2}}$ is bounded as $t \to \infty$. By conservation of energy, this implies that $\|u(t)\|_{H^1}$ is bounded as $t \to \infty$. Thus $\|\mathcal{T}(-t)u(t)\|_{H^1}$ is bounded, which implies that $u^+ \in H^1(\mathbb{R}^N)$. Hence (ii).

Next, we prove (iii). When $\gamma \ge 0$ and $\alpha > 4/N$, the result follows from [6]; and when $\alpha \ge 4/N$ with $\|u(t_0)\|_{H^1}$ small, the result follows from [8]. Regardless of the sign of γ, it is known that if $u \in C((0,\infty), H^1(\mathbb{R}^N))$ is a solution of (1.1) such that $\|u(t)\|_{L^{\alpha+2}} \to 0$ as $t \to \infty$ and if $4/N < \alpha < 4/(N-2)$ (or if $\alpha = 4/N$ and $\|u(t_0)\|_{H^1}$ is small for some $t_0 \ge 0$), then $\mathcal{T}(-t)u(t)$ has a limit in $H^1(\mathbb{R}^N)$ as $t \to \infty$. We give the proof of this last property for completeness. We first show that

$$u \in L^q((1,\infty), W^{1,r}(\mathbb{R}^N)) \tag{3.15}$$

with $r = \alpha + 2$ and $q = 4(\alpha+2)/N\alpha$. First observe that $u \in L^q_{\text{loc}}((0,\infty), W^{1,r}(\mathbb{R}^N))$ (see [1, Theorem 5.3.1 and Remark 5.3.5]). Fix now $S > 0$ to be chosen large enough. It follows from (3.1) that

$$u(t+S) = \mathcal{T}(t)u(S) - i\gamma \int_0^t \mathcal{T}(t-s)(|u|^\alpha u)(s+S)\, ds.$$

It follows from Strichartz' estimate (see, e.g., [1, Theorem 3.2.5]) that

$$\|u\|_{L^q((S,T),W^{1,r})} \le C\|u(S)\|_{H^1} + C\| \, |u|^\alpha u\|_{L^{q'}((S,T),W^{1,r'})}$$

for all $S < T < \infty$, where C is independent of S and T. On the other hand, it follows from Hölder's inequality that

$$\| \, |u|^\alpha u\|_{L^{q'}((S,T),W^{1,r'})} \le C(\sup_{t\ge S} \|u(t)\|_{L^{\alpha+2}})^{\left(1-\frac{4}{N\alpha}\right)(\alpha+2)} \|u\|^{1+\frac{2(4-(N-2)\alpha)}{N\alpha}}_{L^q((S,T),W^{1,r})}. \tag{3.16}$$

Since by conservation of energy $\|u(S)\|_{H^1}^2 \to \|u(t_0)\|_{L^2}^2 + 2E(u(t_0))$ as $S \to \infty$, we deduce from the above inequalities that there exists C, independent of S large and $T > S$, such that

$$\begin{aligned}\|u\|_{L^q((S,T),W^{1,r})} \le C(\|u(t_0)\|_{L^2}^2 + 2E(u(t_0)))^{\frac12}& \\ +C(\sup_{t\ge S} \|u(t)\|_{L^{\alpha+2}})^{\left(1-\frac{4}{N\alpha}\right)(\alpha+2)} &\|u\|^{1+\frac{2(4-(N-2)\alpha)}{N\alpha}}_{L^q((S,T),W^{1,r})}.\end{aligned}$$

It follows, either by choosing S large enough if $\alpha > 4/N$, or by assuming $\|u(t_0)\|_{H^1}$ small enough if $\alpha = 4/N$, that

$$\|u\|_{L^q((S,T),W^{1,r})} \le C,$$

independent of $T > S$. We obtain (3.15) by letting $T \to \infty$. We now set $v(t) = \mathcal{T}(-t)u(t)$ and we observe that

$$v(\tau) - v(t) = -i\gamma \int_t^\tau \mathcal{T}(-s)(|u|^\alpha u)(s)\, ds$$

for all $0 < t < \tau < \infty$. Applying Strichartz' estimate, (3.16), and (3.15), we obtain that

$$\begin{aligned}\|v(\tau) - v(t)\|_{H^1} &\le C\| \, |u|^\alpha u\|_{L^{q'}((t,\tau),W^{1,r'})} \\ &\le C(\sup_{s\ge t} \|u(s)\|_{L^{\alpha+2}})^{\left(1-\frac{4}{N\alpha}\right)(\alpha+2)} \|u\|^{1+\frac{2(4-(N-2)\alpha)}{N\alpha}}_{L^q((t,\infty),W^{1,r})} \\ &\le Ct^{-\beta(\alpha+2)\frac{N\alpha-4}{N\alpha}},\end{aligned}$$

and (iii) follows. (If $\alpha = 4/N$, note that $\|u\|_{L^q((t,\infty),W^{1,r})} \to 0$ as $t \to \infty$.)

Finally, we prove (iv). The proof follows the same outline as for part (iii), except that instead of (3.16) we use the inequality

$$\| \, |u|^\alpha u\|_{L^{q'}((S,T),W^{1,r'})} \le C\|u\|^\alpha_{L^{\frac{1}{\beta}}((S,T),L^{\alpha+2})} \|u\|_{L^q((S,T),W^{1,r})}, \tag{3.17}$$

and we observe that

$$\|u\|_{L^{\frac{1}{\beta}}((S,T),L^{\alpha+2})} \le \Big(\frac{\beta}{\mu-\beta}\Big)^\beta S^{-(\mu-\beta)} \|u\|_{\mathcal{X}_{\alpha,\mu}}. \tag{3.18}$$

This completes the proof. □

Remark 3.11. As is well known, an H^1 scattering theory (global when $\gamma \ge 0$ and "low energy" when $\gamma < 0$) was developed by Ginibre and Velo [6].

Remark 3.12. The previous proof shows that (3.15) under the hypotheses of Proposition 3.10 (iii). It follows that $\liminf_{t\to\infty} t^{\frac{N\alpha}{4(\alpha+2)}}\|u(t)\|_{L^{\alpha+2}} = 0$, and in particular $\liminf_{t\to\infty} t^{\beta}\|u(t)\|_{L^{\alpha+2}} = 0$.

Proposition 3.13. *Suppose* $4/N < \alpha < 4/(N-2)$ *and let* $u \in X_\alpha$ *be a solution of* (1.1) *with initial value* φ. *If* $u(t) \in H^1(\mathbb{R}^N)$ *for all* $t > 0$ *and* $\varphi = 0$, *then* $u(t) \equiv 0$.

Proof. By standard interpolation,

$$\|u(t)\|_{L^2} \le C\|u(t)\|^{\theta}_{H^{-\frac{N\alpha}{2(\alpha+2)}}}\|u(t)\|^{1-\theta}_{H^1}$$

with $\theta\frac{N\alpha}{2(\alpha+2)} = 1-\theta$. Using formula (3.3) and conservation of charge and energy, we deduce that

$$\begin{aligned}\|u(t)\|_{L^2} &\le Ct^{\theta(1-\beta(\alpha+1))} + Ct^{\theta(1-\beta(\alpha+1))}\|u(t)\|^{(1-\theta)\frac{\alpha+2}{2}}_{L^{\alpha+2}}\\ &\le Ct^{\theta(1-\beta(\alpha+1))} + Ct^{\theta(1-\beta(\alpha+1))-\beta(1-\theta)\frac{\alpha+2}{2}}.\end{aligned}$$

Since $\alpha > 4/N$, the powers of t are positive, which contradicts the conservation of charge as $t \downarrow 0$. □

4. Pseudoconformally equivalent solutions. In this paragraph, we consider solutions v of a nonautonomous nonlinear Schrödinger equation obtained from (1.1) by the pseudoconformal transformation. In particular, we give an explicit relationship between the scattering state u^+ and the initial value $v(0)$.

For the rest of this section, we assume that $u, v \in L^\infty_{\text{loc}}((0,\infty), L^{\alpha+2}(\mathbb{R}^N))$ are related by the following formula:

$$v(s,y) = s^{-\frac{N}{2}}e^{i\frac{|y|^2}{4s}}\overline{u}\Big(\frac{1}{s},\frac{y}{s}\Big) = t^{\frac{N}{2}}e^{i\frac{|x|^2}{4t}}\overline{u}(t,x), \tag{4.1}$$

where $s, t > 0$, $x, y \in \mathbb{R}^N$ with $s = 1/t$ and $x = y/s = ty$. It follows (see Lemma 5.9 of [4]) that u is a solution of (1.1) in the sense of Definition 3.1 if and only if v is a solution of

$$v(s) = \mathcal{T}(s-\sigma)v(\sigma) - i\gamma\int_\sigma^s \mathcal{T}(s-\theta)\theta^{\frac{N\alpha-4}{2}}|v(\theta)|^\alpha v(\theta)\,d\theta \tag{4.2}$$

for all $0 < \sigma \le s < \infty$. Moreover, an elementary calculation shows that

$$t^{\beta}\|u(t)\|_{L^{\alpha+2}} = s^{\delta}\|v(s)\|_{L^{\alpha+2}}, \tag{4.3}$$

where

$$\delta = \frac{N\alpha^2 + (N-2)\alpha - 4}{2\alpha(\alpha+2)}. \tag{4.4}$$

In particular, $u \in X_\alpha$ if and only if $v \in \mathcal{X}_{\alpha,\delta}$. More generally,

$$t^{\mu}\|u(t)\|_{L^{\alpha+2}} = s^{\frac{N\alpha}{2(\alpha+2)}-\mu}\|v(s)\|_{L^{\alpha+2}} \tag{4.5}$$

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

Proposition 4.1. *Suppose* (1.4). *Let* $v \in \mathcal{X}_{\alpha,\delta}$ *be a solution of* (4.2). *It follows that there exists* $\psi \in \mathcal{W}_{\alpha,\delta}$ *such that* $v(s) \to \psi$ *in* $\mathcal{S}'(\mathbb{R}^N)$ *as* $s \downarrow 0$ *and* v *satisfies*

$$v(s) = \mathcal{T}(s)\psi - i\gamma\int_0^s \mathcal{T}(s-\theta)\theta^{\frac{N\alpha-4}{2}}|v(\theta)|^\alpha v(\theta)\,d\theta \tag{4.6}$$

for all $s > 0$. *Moreover,*

$$\mathcal{F}\psi = i^{\frac{N}{2}} D_{4\pi}\overline{u^+}, \tag{4.7}$$

where $\mathcal{F}$ *denotes the Fourier transform*

$$\mathcal{F}\phi(x) = \int_{\mathbb{R}^N} e^{-2\pi i x\cdot y}\phi(y)\,dy,$$

and D_μ *denotes the dilation operator defined by* $D_\mu\phi(x) = \mu^{\frac{N}{2}}\phi(\mu x)$.

Proof. The first part of the proposition is proved in the same way as statement (iii) in Proposition 3.3. The second statement is a consequence of Lemma 4.2 below. Indeed, note that formula (4.1) can be written as $\overline{u}(t) = M_{-s}D_s v(s)$. Lemma 4.2 therefore implies that

$$\overline{\mathcal{T}(-t)u(t)} = \mathcal{T}(t)\overline{u}(t) = \mathcal{T}\Big(\frac{1}{s}\Big)M_{-s}D_s v(s) = i^{-\frac{N}{2}}D_{\frac{1}{4\pi}}\mathcal{F}\mathcal{T}(-s)v(s).$$

The result follows by letting $s \downarrow 0$ and thus $t \to \infty$. Indeed, $\mathcal{T}(-s)v(s) \to \psi$ in $\mathcal{S}'(\mathbb{R}^N)$ as $s \downarrow 0$ by construction of ψ. □

Lemma 4.2. *For all* $s > 0$,

$$T\Big(\frac{1}{s}\Big)M_{-s}D_s\mathcal{T}(s) = i^{-\frac{N}{2}}D_{\frac{1}{4\pi}}\mathcal{F},$$

where the multiplication operator M_s *is defined by* $M_s\phi(x) = e^{i\frac{s|x|^2}{4}}\phi(x)$.

Proof. Starting with formula (3.5) in [2] we derive that

$$\begin{aligned}\mathcal{T}\Big(\frac{1}{s}\Big)M_{-s}D_s\mathcal{T}(s) &= i^{-\frac{N}{2}}M_s D_{\frac{s}{4\pi}}\mathcal{F}D_s\mathcal{T}(s) = i^{-\frac{N}{2}}M_s D_{\frac{s}{4\pi}}D_{\frac{1}{s}}\mathcal{F}\mathcal{T}(s)\\ &= i^{-\frac{N}{2}}M_s D_{\frac{1}{4\pi}}M_{-16\pi^2 s}\mathcal{F} = i^{-\frac{N}{2}}D_{\frac{1}{4\pi}}\mathcal{F}.\end{aligned}$$

This completes the proof. □

Remark 4.3. Formula (4.7) is analogous to formula (3.11) in [2] as well as the last formula in Proposition 3.14 in [3]. The difference between this paper and the papers [2], [3] is the specific form of the pseudoconformal transformation used. In [2], [3] the transformation fixed $t = 0$ and brought ∞ to $t = 1$. In the present paper, the transformation we use exchanges $t = \infty$ and $t = 0$. Moreover, the class of solutions considered in the present paper is different.

5. Wave operators and their properties. In this section, we prove that, given a scattering state u^+ sufficiently small in the space W_α, there is a unique solution u of (1.1) sufficiently small in X_α which has u^+ as a scattering state. The construction is similar to Theorem 2.1 in [4] where we prove that sufficiently small initial values $\varphi \in W_\alpha$ give rise to solutions of (1.1) in X_α. As a result, we will have constructed a "low energy" scattering theory for (1.1) in the space W_α.

Theorem 5.1. *Suppose* (1.2) *and* (1.4). *Assume* $\rho, M > 0$ *satisfy the inequality*

$$\rho + KM^{\alpha+1} \le M,$$

where $K = K(\alpha,\gamma)$ *is given by* (5.5) *below. Let* $u^+ \in W_\alpha$ *be such that* $\|u^+\|_{W_\alpha} \le \rho$. *It follows that there exists a unique solution* $u \in X_\alpha$ *of* (1.1) *in the sense of Definition* 3.1

such that $\|u\|_{X_\alpha} \le M$ *and* $\mathcal{T}(-t)u(t) \to u^+$ *in* $\mathcal{S}'(\mathbb{R}^N)$ *as* $t \to \infty$*. Furthermore, the following properties hold:*

(i) *Let* $u^+, v^+ \in W_\alpha$ *with* $\|u^+\|_{W_\alpha}, \|v^+\|_{W_\alpha} \le \rho$ *and let* u *and* v *be the resulting solutions of* (1.1)*. It follows that* $\|u - v\|_{X_\alpha} \le C\|u^+ - v^+\|_{W_\alpha}$*, for some constant* C *independent of* u^+ *and* v^+*.*

(ii) *Let* $u^+, v^+ \in W_\alpha$ *with* $\|u^+\|_{W_\alpha}, \|v^+\|_{W_\alpha} \le \rho$ *and let* u *and* v *be the resulting solutions of* (1.1)*. Given* $\mu > 0$*, there exists* $\rho_1 = \rho_1(\alpha, \gamma, \mu) > 0$ *such that if* $\|u^+\|_{W_\alpha} + \|u^+\|_{W_\alpha} \le \rho_1$ *and* $u^+ - v^+ \in \mathcal{W}_{\alpha,\mu}$*, then* $u - v \in \mathcal{X}_{\alpha,\mu}$*.*

(iii) *Let* $u^+ \in W_\alpha$ *with* $\|u^+\|_{W_\alpha} \le \rho$ *and let* u *be the resulting solution of* (1.1)*. Given* $\mu > 0$*, there exists* $\rho_2 = \rho_2(\alpha, \gamma, \mu) > 0$ *such that if* $\|u^+\|_{W_\alpha} \le \rho_2$ *and* $u^+ \in \mathcal{W}_{\alpha,\mu}$*, then* $u \in \mathcal{X}_{\alpha,\mu}$*.*

Proof. The basic idea of the proof is to use a contraction mapping argument to prove global existence of solutions of the equation

$$u(t) = \mathcal{T}(t)u^+ + i\gamma \int_t^\infty \mathcal{T}(t-s)|u(s)|^\alpha u(s)\, ds \tag{5.1}$$

for all $t > 0$, which is equation (3.4) with $\tau = t$.

For $u^+ \in W_\alpha$ and $u \in X_\alpha$, we set

$$(\mathcal{Q}_{u^+}u)(t) = \mathcal{T}(t)u^+ + i\gamma \int_t^\infty \mathcal{T}(t-s)|u(s)|^\alpha u(s)\, ds \tag{5.2}$$

for all $t > 0$. In other words, $(\mathcal{Q}_{u^+}u)(t) = \mathcal{T}(t)u^+ + i\gamma\mathcal{G}(t, \infty, t, u)$. This definition makes sense by Proposition 2.4. In particular, it follows from (2.18) applied with $v = 0$ and $\mu = \beta$ that $\mathcal{Q}_{u^+}u \in X_\alpha$ and

$$\|\mathcal{Q}_{u^+}u\|_{X_\alpha} \le \|u^+\|_{W_\alpha} + C_1\|u\|_{X_\alpha}^{\alpha+1}. \tag{5.3}$$

Moreover, if $u^+, v^+ \in W_\alpha$ and $u, v \in X_\alpha$, then it follows also from (2.18) that

$$\|\mathcal{Q}_{u^+}u - \mathcal{Q}_{v^+}v\|_{X_\alpha} \le \|u^+ - v^+\|_{W_\alpha} + C_2 \max\{\|u\|_{X_\alpha}, \|v\|_{X_\alpha}\}^\alpha \|u - v\|_{X_\alpha}. \tag{5.4}$$

It is clear now that if $M, \rho > 0$ satisfy $\rho + KM^{\alpha+1} \le M$ with

$$K = K(\alpha, \gamma) = \max\{C_1, C_2\}, \tag{5.5}$$

then given $u^+ \in W_\alpha$ with $\|u^+\|_{W_\alpha} \le \rho$, $\mathcal{Q}_{u^+}$ is a strict contraction on the set $B_M = \{u \in X_\alpha;\ \|u\|_{X_\alpha} \le M\}$ equipped with the distance induced by the norm in X_α. Thus $\mathcal{Q}_{u^+}$ has a unique fixed point, which is the unique solution of (5.1) in B_M. We also obtain the estimate of property (i).

Next we show that u is a solution of (1.1) in the sense of Definition 3.1. We observe first that by (2.17), we may commute $\mathcal{T}(t)$ with the integral in (5.1); and so, u satisfies

$$\mathcal{T}(-t)u(t) = u^+ + i\gamma \int_t^\infty \mathcal{T}(-s)|u(s)|^\alpha u(s)\, ds. \tag{5.6}$$

The integral converges strongly in $L^{\alpha+2}(\mathbb{R}^N)$ and it follows in particular that

$$\mathcal{T}(-t)u(t) - u^+ \underset{t\to\infty}{\longrightarrow} 0 \tag{5.7}$$

in $L^{\alpha+2}(\mathbb{R}^N)$. Furthermore, (5.6) implies that

$$\mathcal{T}(-t)u(t) - \mathcal{T}(-\tau)u(\tau) = i\gamma \int_t^\tau \mathcal{T}(-s)|u(s)|^\alpha u(s)\,ds$$

for all (finite) $t, \tau > 0$, where now the integral converges in both $L^{\alpha+2}(\mathbb{R}^N)$ and $H^{-\frac{N\alpha}{2(\alpha+2)}}$. By applying $\mathcal{T}(t)$, we see that u is a solution of (1.1) in the sense of Definition 3.1. This completes the proof of the main statement of the theorem and property (i).

Finally, we prove property (ii) (property (iii) being a special case). If $u^+, v^+ \in W_\alpha$ with $u^+ - v^+ \in \mathcal{W}_{\alpha,\mu}$ and if $u, v \in X_\alpha$ with $u - v \in \mathcal{X}_{\alpha,\mu}$, then it follows from (2.19) that $\mathcal{Q}_{u^+}u - \mathcal{Q}_{v^+}v \in \mathcal{X}_{\alpha,\mu}$ and

$$(5.8)\quad \|\mathcal{Q}_{u^+}u - \mathcal{Q}_{v^+}v\|_{\mathcal{X}_{\alpha,\mu}} \le \|u^+ - v^+\|_{\mathcal{W}_{\alpha,\mu}} + C_3 \max\{\|u\|_{X_\alpha}, \|v\|_{X_\alpha}\}^\alpha \|u - v\|_{\mathcal{X}_{\alpha,\mu}}$$

with $C_3 = C_3(\alpha, \gamma, \mu)$. To prove the statement, we fix $0 < \rho_1 \le \rho$ and $0 < M_1 \le M$ small enough so that

$$\rho_1 + K(\alpha, \gamma) M_1^{\alpha+1} \le M_1 \quad \text{and} \quad C_3 M_1^\alpha < 1.$$

Assuming $\|v^+\|_{W_\alpha} \le \rho_1$, it follows that the corresponding solution v satisfies $\|v\|_{X_\alpha} \le M_1$. Next, assuming also that $\|u^+\|_{W_\alpha} \le \rho_1$, we carry out a new contraction mapping argument in the set

$$\widetilde{B} = \{w \in X_\alpha;\, w - v \in \mathcal{X}_{\alpha,\mu}, \|w\|_{X_\alpha} \le M_1, \|w - v\|_{\mathcal{X}_{\alpha,\mu}} \le L_1\}$$

with

$$L_1 = \frac{\|u^+ - v^+\|_{\mathcal{W}_{\alpha,\mu}}}{1 - C_3 M_1^\alpha}.$$

$\widetilde{B}$ is a complete metric space using the distance induced by the norm in X_α. The only new element in the contraction mapping argument is to show that $\|\mathcal{Q}_{u^+}w - v\|_{\mathcal{X}_{\alpha,\mu}} \le L_1$ for all $w \in \widetilde{B}$. This is an immediate consequence of (5.8) and the definition of L_1, and the fact that $\mathcal{Q}_{v^+}v = v$. The resulting solution u of (1.1) is clearly the same as obtained in B_M since $\widetilde{B} \subset B_M$. □

Remark 5.2. In fact, the proof of Theorem 5.1 never uses the fact that $\alpha > \alpha_0$, which is the same as the condition $\beta(\alpha + 1) < 1$. Indeed, this last condition is needed only when the integrand in (3.1) needs to be estimated as $s \downarrow 0$. On the other hand, the theorem is vacuous if $\alpha < \alpha_0$ since $W_\alpha = \{0\}$.

Proposition 5.3. *Suppose* (1.2) *and* (1.4). *Let* $u^+ \in W_\alpha$ *be such that* $\|u^+\|_{W_\alpha} \le \rho$ *as in Theorem* 5.1 *and let* u *be the resulting solution of* (1.1). *If, in addition,* $u^+ \in H^1(\mathbb{R}^N)$ *and* $\alpha \ge 4/N$, *then* $u(t) \in H^1(\mathbb{R}^N)$ *for all* $t > 0$.

Proof. Let ρ, M be as in Theorem 5.1 and let u^+ and u be as above. Let $r = \alpha + 2$ and $q = 4(\alpha + 2)/N\alpha$ and observe that by Strichartz' estimate $\mathcal{T}(\cdot)u^+ \in L^q(\mathbb{R}, W^{1,r}(\mathbb{R}^N))$. Given $T > 0$, consider the set

$$B_M^T = \Big\{u \in L^\infty((T,\infty), L^{\alpha+2}(\mathbb{R}^N)) \cap L^q((T,\infty), W^{1,r}(\mathbb{R}^N));$$
$$\sup_{t \ge T} t^\beta \|u(t)\|_{L^{\alpha+2}} \le M, \|u\|_{L^q((T,\infty),W^{1,r})} \le 2\|\mathcal{T}(\cdot)u^+\|_{L^q((T,\infty),W^{1,r})}\Big\}.$$

It is clear that B_M^T equipped with the distance $d(u,v) = \sup_{t\ge T} t^\beta \|u(t) - v(t)\|_{L^{\alpha+2}}$ is a complete metric space. On the other hand, it follows from (5.3), (5.4), and (5.5) that (with $\mathcal{Q}$ defined by (5.2))

$$\sup_{t\ge T} t^\beta \|\mathcal{Q}_{u^+} u\|_{L^{\alpha+2}} \le \rho + KM^{\alpha+1} \le M, \tag{5.9}$$

$$d(\mathcal{Q}_{u^+} u, \mathcal{Q}_{u^+} v) \le KM^\alpha d(u,v) \tag{5.10}$$

for all $u, v \in B_M^T$. Furthermore, it follows from Strichartz' estimate (see, e.g., [1, Theorem 3.2.5]) that

$$\begin{aligned}\|\mathcal{Q}_{u^+} u\|_{L^q((T,\infty),W^{1,r})} &\le \|\mathcal{T}(\cdot)u^+\|_{L^q((T,\infty),W^{1,r})} + C\| \, |u|^\alpha u\|_{L^{q'}((T,\infty),W^{1,r'})} \\ &\le \|\mathcal{T}(\cdot)u^+\|_{L^q((T,\infty),W^{1,r})} + C\left(\sup_{t\ge T}\|u(t)\|_{L^{\alpha+2}}\right)^{\left(1-\frac{4}{N\alpha}\right)(\alpha+2)} \|u\|_{L^q((T,\infty),W^{1,r})}^{1+\frac{2(4-(N-2)\alpha)}{N\alpha}},\end{aligned}$$

where the last inequality follows from (3.16). Therefore,

$$\begin{aligned}\|\mathcal{Q}_{u^+} u\|_{L^q((T,\infty),W^{1,r})} \le{}& \|\mathcal{T}(\cdot)u^+\|_{L^q((T,\infty),W^{1,r})} \\ &+CT^{-\beta\left(1-\frac{4}{N\alpha}\right)(\alpha+2)}(2\|\mathcal{T}(\cdot)u^+\|_{L^q((T,\infty),W^{1,r})})^{1+\frac{2(4-(N-2)\alpha)}{N\alpha}} \\ \le{}& 2\|\mathcal{T}(\cdot)u^+\|_{L^q((T,\infty),W^{1,r})}\end{aligned} \tag{5.11}$$

if T is large enough. (5.9), (5.10), and (5.11) imply that $\mathcal{Q}_{u^+}$ has a unique fixed point in B_M^T, which clearly coincides with the solution constructed in Theorem 5.1 by (5.10). Thus $u \in L^q((T,\infty), W^{1,r}(\mathbb{R}^N))$ for T large enough. Applying again Strichartz' inequality, we now see that $u \in C([T,\infty), H^1(\mathbb{R}^N))$. The result now follows from Proposition 3.10 (i). □

PROPOSITION 5.4. *Suppose* (1.2) *and* (1.4). *Let* $u^+ \in W_\alpha \cap \mathcal{W}_{\alpha,\mu}$ *for some* $\mu > \beta$. *Suppose* $\|u^+\|_{W_\alpha} \le \min\{\rho, \rho_1\}$ *as in Theorem* 5.1 *parts* (i) *and* (iii) *and let u be the resulting solution of* (1.1). *If, in addition,* $u^+ \in H^1(\mathbb{R}^N)$, *then* $u(t) \in H^1(\mathbb{R}^N)$ *for all* $t > 0$.

Proof. The proof is similar to the proof of Proposition 5.3 above, except that instead of the inequality (3.16), we use the inequalities (3.17) and (3.18). □

PROPOSITION 5.5. *Suppose* (1.2) *and* (1.4). *Let* $u^+ \in H^1(\mathbb{R}^N) \cap \mathcal{W}_{\alpha,\mu}$ *for some* $\mu > \beta$. *Suppose* $\gamma > 0$ *or* $\alpha < 4/N$. *Then there exists a unique solution* $u \in C([0,\infty), H^1(\mathbb{R}^N)) \cap \mathcal{X}_{\alpha,\mu}$ *of* (1.1) *whose scattering state is* u^+.

Proof. The proof is again similar to the proof of Proposition 5.3 above, and we use the same notation. Fix $M \ge 2\sup_{t>0} t^\mu \|\mathcal{T}(t)u^+\|_{L^{\alpha+2}}$, $M \ge 2\|\mathcal{T}(\cdot)u^+\|_{L^q(\mathbb{R},W^{1,r})}$. Given $T, M > 0$, consider the set

$$B_M^T = \left\{ u \in L^\infty((T,\infty), L^{\alpha+2}(\mathbb{R}^N)) \cap L^q((T,\infty), W^{1,r}(\mathbb{R}^N)); \right. \\ \left. \sup_{t\ge T} t^\mu \|u(t)\|_{L^{\alpha+2}} \le M, \|u\|_{L^q((T,\infty),W^{1,r})} \le M \right\}.$$

It is clear that B_M^T equipped with the distance $d(u,v) = \sup_{t\ge T} t^\mu \|u(t) - v(t)\|_{L^{\alpha+2}}$ is a complete metric space. Using (3.17) and (3.18) instead of (3.16), we still obtain the final inequality in (5.11), i.e., if T is large enough, then

$$\|\mathcal{Q}_{u^+} u\|_{L^q((T,\infty),W^{1,r})} \le 2\|\mathcal{T}(\cdot)u^+\|_{L^q(\mathbb{R},W^{1,r})} \le M. \tag{5.12}$$

Next,

$$\sup_{t\geq T} t^{\mu}\|\mathcal{Q}_{u^+}\|_{L^{\alpha+2}} \leq \sup_{t\geq T} t^{\mu}\|\mathcal{T}(t)u^+\|_{L^{\alpha+2}} + \gamma M^{\alpha+1}\sup_{t\geq T} t^{\mu}\int_t^{\infty}|t-s|^{-\frac{N\alpha}{2(\alpha+2)}}s^{-\mu(\alpha+1)}\,ds \leq \sup_{t\geq T} t^{\mu}\|\mathcal{T}(t)u^+\|_{L^{\alpha+2}} + C\gamma M^{\alpha+1}T^{1-\frac{N\alpha}{2(\alpha+2)}-\alpha\mu}.$$

Since $1-\frac{N\alpha}{2(\alpha+2)}-\alpha\mu < 1-\frac{N\alpha}{2(\alpha+2)}-\alpha\beta = 0$,

$$\sup_{t\geq T} t^{\mu}\|\mathcal{Q}_{u^+}\|_{L^{\alpha+2}} \leq M \tag{5.13}$$

if T is large enough. In a similar way, we see that

$$d(\mathcal{Q}_{u^+}(u),\mathcal{Q}_{u^+}(v)) \leq C\gamma M^{\alpha}T^{1-\frac{N\alpha}{2(\alpha+2)}-\alpha\mu}d(u,v) \leq \frac{1}{2}d(u,v) \tag{5.14}$$

if T is large enough. By (5.12), (5.13), and (5.14), $\mathcal{Q}_{u^+}$ has a unique fixed point $u \in B_T^M$ for T large enough. By Strichartz' inequality, we see that $u \in C([T,\infty),H^1(\mathbb{R}^N))$. Since $\gamma > 0$ or $\alpha < 4/N$, all H^1 solutions of (1.1) are global, so that u can be extended to a solution $u \in C([0,\infty),H^1(\mathbb{R}^N))$. Since $u \in B_M^T$, and $H^1(\mathbb{R}^N) \hookrightarrow L^{\alpha+2}(\mathbb{R}^N)$, we clearly have $u \in \mathcal{X}_{\alpha,\mu}\cap X_{\alpha}$. We conclude as in Theorem 5.1 that u^+ is the scattering state of u.

Finally we show uniqueness. Consider two solutions $u,v \in C([0,\infty),H^1(\mathbb{R}^N))\cap \mathcal{X}_{\alpha,\mu}$ of (1.1) with the same scattering state $u^+ \in H^1(\mathbb{R}^N)\cap\mathcal{W}_{\alpha,\mu}$. It follows (see the proof of Proposition 3.10 (iv)) that $u,v \in L^q((0,\infty),W^{1,r}(\mathbb{R}^N))$. Consider now the set B_M^T with M sufficiently large so that

$$M \geq \max\{\|u\|_{\mathcal{X}_{\alpha,\mu}},\|v\|_{\mathcal{X}_{\alpha,\mu}},\|u\|_{L^q((0,\infty),W^{1,r})},\|v\|_{L^q((0,\infty),W^{1,r})}\}.$$

In particular, $u,v \in B_M^T$ for all $T>0$. Applying (5.14) with this choice of M, we see that $u(t)=v(t)$ for t sufficiently large. By uniqueness of H^1 solutions, we conclude that $u=v$. □

6. Self-similar solutions. Let $p \in \mathbb{C}$ with $\operatorname{Re} p = 2/\alpha$. Given $\lambda > 0$ and $u \in \mathcal{S}'((0,\infty)\times\mathbb{R}^N)$, let $u_\lambda(t,x) = \lambda^p u(\lambda^2 t,\lambda x)$. (Even if u is not a function, this is obviously defined by duality.) If $u \in X_\alpha$, then it follows that $u_\lambda \in X_\alpha$ and $\|u_\lambda\|_{X_\alpha} = \|u\|_{X_\alpha}$. Moreover, if $u \in X_\alpha$ is a solution of (1.1), then u_λ is also a solution. In addition, if φ and u^+ are the initial values and scattering states of u and φ_λ and u_λ^+ are the initial values and scattering states of u_λ, then $\varphi_\lambda(x) = \lambda^p\varphi(\lambda x)$ and $u_\lambda^+(x) = \lambda^p u^+(\lambda x)$.

We recall that a solution $u \in X_\alpha$ of (1.1) is self-similar if $u_\lambda = u$ for all $\lambda > 0$.

PROPOSITION 6.1. *Suppose* (1.4) *and let* $p \in \mathbb{C}$ *with* $\operatorname{Re} p = 2/\alpha$.

(i) *If* $u \in X_\alpha$ *is a self-similar solution of* (1.1), *it follows that the initial value* φ *and the scattering state* u^+ *are homogeneous tempered distributions of degree* $-p$.

(ii) *Let* $u^+ \in W_\alpha$ *be such that* $\|u^+\|_{W_\alpha} \leq \rho$ *as in Theorem* 5.1. *Let* u *be the resulting solution of* (1.1) *with scattering state* u^+ *as provided by Theorem* 5.1. *If in addition* u^+ *is a homogeneous tempered distribution of degree* $-p$, *then* u *is a self-similar solution of* (1.1).

Proof. (i) follows from formula (3.4) with $\tau = 0$ and from formula (3.8) by a straightforward calculation using $u = u_\lambda$ and the commutation relation $(\mathcal{T}(t)\psi)_\lambda = \mathcal{T}(t/\lambda^2)(\psi_\lambda)$ for all $t \in \mathbb{R}$ and $\lambda > 0$.

The proof of statement (ii) is the same as the proof of the analogous statement in Proposition 4.3 of [4] for initial values. For each $\lambda > 0$, u_λ and u are both solutions of (1.1) with $\|u_\lambda\|_{X_\alpha} = \|u\|_{X_\alpha} \le M$ as in Theorem 5.1 with the same scattering state. By the uniqueness part of Theorem 5.1, $u_\lambda = u$. □

Remark 6.2. Let $f \in L^1_{\mathrm{loc}}(\mathbb{R}^N)$, let $p \in \mathbb{C}$ with $\operatorname{Re} p = 2/\alpha$, and set $u(t,x) = t^{-\frac{p}{2}} f(x/\sqrt{t})$ for $t > 0$ and $x \in \mathbb{R}^N$. It follows that $f \in L^{\alpha+2}(\mathbb{R}^N)$ if and only if $u \in X_\alpha$.

Proposition 6.3. *In addition to* (1.4)*, suppose $N \ge 3$. Let $p \in \mathbb{C}$ with* $\operatorname{Re} p = 2/\alpha$. *Let $u \in X_\alpha$ be a self-similar solution of* (1.1) *with the profile f, i.e., $u(t,x) = t^{-\frac{p}{2}} f(x/\sqrt{t})$. If $\nabla f \in L^2(\mathbb{R}^N)$, then $u^+ = 0$.*

Proof. We first show that

$$\mathcal{T}(-t)u(t) \in L^{\frac{2N}{N-2}}(\mathbb{R}^N) \tag{6.1}$$

for all $t > 0$. Set $w = \mathcal{T}(-t)u(t)$. Since $w \in W_\alpha$ by Proposition 3.3, it follows that for $\tau > 0$,

$$\mathcal{T}(\tau)w \in L^{\alpha+2}(\mathbb{R}^N). \tag{6.2}$$

Since $\nabla f \in L^2(\mathbb{R}^N)$, it is clear that $\nabla u(t) \in L^2(\mathbb{R}^N)$. Thus

$$\nabla \mathcal{T}(\tau)w \in L^2(\mathbb{R}^N). \tag{6.3}$$

It follows from (6.2), (6.3), and an obvious truncation argument that

$$\|\mathcal{T}(\tau)w\|_{L^{\frac{2N}{N-2}}} \le C\|\nabla \mathcal{T}(\tau)w\|_{L^2} = C\|\nabla w\|_{L^2}.$$

We obtain (6.1) by letting $\tau \downarrow 0$.

It therefore follows from (3.4) with $\tau = 0$, from (6.1), and from (3.9) that $u^+ \in L^{\frac{2N}{N-2}}(\mathbb{R}^N) + L^{\alpha+2}(\mathbb{R}^N)$. Since u^+ is homogeneous of degree $-p$ by Proposition 6.1, and since $N\alpha/2 < \alpha + 2 < 2N/(N-2)$, u^+ must be 0, as shown by Lemma 6.4 below. □

Lemma 6.4. *Let $1 \le r \le \infty$ and let $v \in L^r_{\mathrm{loc}}(\mathbb{R}^N)$. If v is homogeneous of degree $-k + i\omega$ with $k \ge N/r$ and $k > 0$ if $r = \infty$, then $v \equiv 0$.*

Proof. Let $0 < \lambda < 1$. Since $\lambda^k |v(\lambda x)| \equiv |v(x)|$, it follows that

$$\|v\|_{L^r(\{|x|<1\})} = \lambda^{k-\frac{N}{r}} \|v\|_{L^r(\{|x|<\lambda\})}.$$

If $r < \infty$, then $\|v\|_{L^r(\{|x|<1\})} \le \|v\|_{L^r(\{|x|<\lambda\})}$, thus $v = 0$ on $\{\lambda < |x| < 1\}$. If $r = \infty$, then $\|v\|_{L^r(\{|x|<1\})} < \|v\|_{L^r(\{|x|<1\})}$, thus $v = 0$ on $\{|x| < 1\}$. By homogeneity, we deduce in both cases that $v \equiv 0$. □

Corollary 6.5. *Under the assumptions of Proposition* 6.3*, let $u \in X_\alpha$ be a self-similar solution of* (1.1) *with the profile $f \in L^{\alpha+2}(\mathbb{R}^N)$, $f \not\equiv 0$. If $\|f\|_{L^{\alpha+2}} \le M$ with M as in the statement of Theorem* 5.1*, then $\nabla f \notin L^2(\mathbb{R}^N)$.*

Proof. If $\nabla f \in L^2(\mathbb{R}^N)$, then by Proposition 6.3 $u^+ = 0$, and by Theorem 5.1 it follows that $u(t) \equiv 0$ (note that $\|u\|_{X_\alpha} = \|f\|_{L^{\alpha+2}}$). □

Remark 6.6. Remark 3.12 shows that if $\alpha = 4/N$, then there is no H^1 self-similar solution of (1.1) with small H^1 norm. (If $N \ge 3$, this is also a consequence of Corollary 6.5.) It follows from [9] that there is no *radially symmetric* H^1 self-similar solution of (1.1).

Remark 6.7. If $u \in X_\alpha$ is a self-similar solution of (1.1) with profile f and if $\mathcal{T}(-1)f \in L^{\frac{N\alpha}{2}}_{\mathrm{loc}}(\mathbb{R}^N)$, then the proof of Proposition 6.3 shows that $u^+ = 0$. This is true in particular if $f \in L^r(\mathbb{R}^N)$ for some $r \le 2$ in the case $\alpha \le 4/N$ and $r \le N\alpha/(N\alpha - 2)$ in the case $\alpha > 4/N$. This is also true if $f \in \dot{H}^s(\mathbb{R}^N)$ for some $s \ge 0$ in the case $\alpha \le 4/N$ and $s \ge N/2 - 2/\alpha$ in the case $\alpha > 4/N$.

We now establish a partial converse of Proposition 6.3.

Proposition 6.8. *In addition to* (1.4), *suppose* $N \ge 3$. *Let* $p \in \mathbb{C}$ *with* $\operatorname{Re} p = 2/\alpha$. *Let* $u \in X_\alpha$ *be a self-similar solution of* (1.1) *with the profile* f, *i.e.,* $u(t,x) = t^{-\frac{p}{2}} f(x/\sqrt{t})$. *Suppose further that* f *is radially symmetric and that* $f \in L^\rho(\mathbb{R}^N)$ *for some* $\alpha + 1 \le \rho < \frac{2N}{N+2}(\alpha+1)$. *If* $u^+ = 0$, *then* $\nabla f \in L^2(\mathbb{R}^N)$.

Remark 6.9. Note that $\frac{N\alpha}{2} < \frac{2N}{N+2}(\alpha+1) < \alpha + 2$, since $\alpha < 4/(N-2)$.

Proof of Proposition 6.8. Set

$$v = \int_1^\infty \mathcal{T}(-s)|u(s)|^\alpha u(s)\, ds.$$

By formula (3.4), it suffices to show that $\nabla v \in L^2(\mathbb{R}^N)$.

Let $g = \mathcal{T}(-1)|f|^\alpha f$. Since $f \in L^{\alpha+2}(\mathbb{R}^N)$, $g \in L^{\alpha+2}(\mathbb{R}^N)$; and since $f \in L^\rho(\mathbb{R}^N)$, $g \in L^{\frac{\rho}{\rho-\alpha-1}}(\mathbb{R}^N)$. (Note that $q := \frac{\rho}{\rho-\alpha-1} > \frac{2N}{N-2}$.) It follows from the scaling properties of $(\mathcal{T}(t))_{t\in\mathbb{R}}$ with respect to dilations that

$$v(r) = \int_1^\infty s^{-1-\frac{p}{2}} g\Big(\frac{r}{\sqrt{s}}\Big)\, ds,$$

where $r = |x|$. Setting $\sigma = r/\sqrt{s}$, this implies

$$v(r) = 2r^{-p} \int_0^r \sigma^{p-1} g(\sigma)\, d\sigma.$$

Thus,

$$\partial_r v = 2r^{-1} g(r) - 2pr^{-p-1} \int_0^r \sigma^{p-1} g(\sigma)\, d\sigma = 2r^{-1} g(r) - h(r).$$

We now estimate the right-hand side in $L^2(\mathbb{R}^N)$.

$$\|r^{-1} g(r)\|_{L^2(\{|x|>1\})} \le \|g\|_{L^{\alpha+2}} \|r^{-1}\|_{L^{\frac{2(\alpha+2)}{\alpha}}(\{|x|>1\})} < \infty,$$

since $\alpha < 4/(N-2)$; and

$$\|r^{-1} g(r)\|_{L^2(\{|x|<1\})} \le \|g\|_{L^q} \|r^{-1}\|_{L^{\frac{2q}{q-2}}(\{|x|<1\})} < \infty,$$

since $q > 2N/(N-2)$. Thus, $r^{-1} g(r) \in L^2(\mathbb{R}^N)$. Next, we have

$$\Big| \int_0^r \sigma^{p-1} g(\sigma)\, d\sigma \Big| \le C r^{\frac{2}{\alpha} - \frac{N}{\alpha+2}} \|g\|_{L^{\alpha+2}}.$$

Since $-\frac{2}{\alpha} - 1 + \frac{2}{\alpha} - \frac{N}{\alpha+2} < -\frac{N}{2}$, it follows that $h \in L^2(\{|x| > 1\})$. A similar calculation where $\|g\|_{L^q}$ is used in place of $\|g\|_{L^{\alpha+2}}$ shows that $h \in L^2(\{|x| < 1\})$. □

7. Scattering theory versus asymptotically self-similar solutions. In this section we exhibit various classes of scattering states u^+ and initial values φ in $W_\alpha \cap H^1(\mathbb{R}^N)$ for which the asymptotic rate of decay of $\|u(t)\|_{L^{\alpha+2}}$ as $t \to \infty$ can be precisely determined. These classes are determined either by the asymptotic behavior of $u^+(x)$ or $\varphi(x)$ as $|x| \to \infty$ or by the asymptotic behavior of $\widehat{u^+}(\xi)$ or $\widehat{\varphi}(\xi)$ as $|\xi| \to 0$.

As usual, we assume (1.4). In addition, we consider $q \in \mathbb{C}$ such that

$$\frac{\alpha+2}{\alpha+1} < \frac{N}{\operatorname{Re} q} < \alpha + 2. \tag{7.1}$$

Note that if $\operatorname{Re} q = 2/\alpha$, then (7.1) is the same as (1.4). However, in what follows we explicitly exclude this possibility by requiring the following additional conditions:

$$\operatorname{Re} q > \frac{2}{\alpha}, \tag{7.2}$$

$$\operatorname{Re} q > \frac{N}{2}. \tag{7.3}$$

Suppose that $\psi(x)$ is C^∞ away from the origin and homogeneous of degree $-q$. (It follows that $\widehat{\psi}(\xi)$ is also C^∞ away from the origin and homogeneous of degree $q - N$. See [15, p. 262].) By (the proof of) Theorem V.1–3 in [12] (see also Proposition 3.9 in [4], Lemma 3.2 in [5], and Theorem 3 in [13]), (7.1) implies that $\mathcal{T}(t)\psi \in L^{\alpha+2}(\mathbb{R}^N)$ for all $t > 0$; and so,

$$t^\nu \|\mathcal{T}(t)\psi\|_{L^{\alpha+2}} = \|\mathcal{T}(1)\psi\|_{L^{\alpha+2}} \tag{7.4}$$

for all $t > 0$ with

$$\nu = \frac{\operatorname{Re} q}{2} - \frac{N}{2(\alpha+2)} = \beta + \frac{\operatorname{Re} q}{2} - \frac{1}{\alpha} > \beta \tag{7.5}$$

by formula (3.6) in [4]. Note that

$$\nu < \frac{N\alpha}{2(\alpha+2)}, \tag{7.6}$$

by (7.1). Note also that $\mathcal{T}(t)\psi$ is a self-similar solution of the linear Schrödinger equation, with a different homogeneity than the self-similar solutions of (1.1) in the previous section. More precisely, $v(t,x) = \mathcal{T}(t)\psi(x)$ satisfies $\lambda^q v(\lambda^2 t, \lambda x) = v(t,x)$ for all $\lambda > 0$ and all $t \geq 0$, $x \in \mathbb{R}^N$.

Let $\eta \in C_{\mathrm{c}}^\infty(\mathbb{R}^N)$ and $\eta(x) \equiv 1$ in a neighborhood of the origin, and set

$$\Psi = (1 - \eta)\psi. \tag{7.7}$$

Note that $\eta\psi = \psi - \Psi \in L^{\frac{\alpha+2}{\alpha+1}}(\mathbb{R}^N)$ by (7.1). Also, we define Θ by

$$\widehat{\Theta} = \eta\widehat{\psi}. \tag{7.8}$$

Note that, by the arguments on pp. 88–89 in [12], we have that $\psi - \Theta \in L^{\frac{\alpha+2}{\alpha+1}}(\mathbb{R}^N)$.

We begin with the following lemma.

Lemma 7.1. *Assume* (1.4), (7.1), (7.2), *and* (7.3) *and let* Ψ *and* Θ *be as above. It follows that* $\Psi, \Theta \in H^1(\mathbb{R}^N) \cap W_\alpha \cap \mathcal{W}_{\alpha,\nu}$ *with* ν *defined by* (7.5).

Proof. We consider first Ψ. It follows from (7.3) and the homogeneity of ψ that $\Psi \in H^1(\mathbb{R}^N)$; and so

$$\sup_{t\geq 0} \|\mathcal{T}(t)\Psi\|_{L^{\alpha+2}} \leq C \sup_{t\geq 0} \|\mathcal{T}(t)\Psi\|_{H^1} \leq C\|\Psi\|_{H^1} < \infty. \tag{7.9}$$

On the other hand, since $\psi - \Psi = \eta\psi \in L^{\frac{\alpha+2}{\alpha+1}}(\mathbb{R}^N)$,

$$\|\mathcal{T}(t)\Psi\|_{L^{\alpha+2}} \leq \|\mathcal{T}(t)(\Psi-\psi)\|_{L^{\alpha+2}} + \|\mathcal{T}(t)\psi\|_{L^{\alpha+2}} \leq Ct^{-\frac{N\alpha}{2(\alpha+2)}} + Ct^{-\nu},$$

where we also used (7.4). Applying (7.6), we deduce that

$$\sup_{t\geq 1} t^\nu \|\mathcal{T}(t)\Psi\|_{L^{\alpha+2}} < \infty. \tag{7.10}$$

The result now follows from (7.5), (7.9), and (7.10).

Next we turn to Θ. By (7.3) and the homogeneity of $\widehat{\psi}$, we have that $\eta\widehat{\psi} \in L^2(\mathbb{R}^N)$ and $\xi\eta(\xi)\widehat{\psi}(\xi) \in L^2(\mathbb{R}^N)$, and so $\Theta \in H^1(\mathbb{R}^N)$. Furthermore, since $\varphi - \Theta \in L^{\frac{\alpha+2}{\alpha+1}}(\mathbb{R}^N)$, it follows that

$$\|\mathcal{T}(t)\Theta\|_{L^{\alpha+2}} \leq \|\mathcal{T}(t)(\Theta-\psi)\|_{L^{\alpha+2}} + \|\mathcal{T}(t)\psi\|_{L^{\alpha+2}} \leq Ct^{-\frac{N\alpha}{2(\alpha+2)}} + Ct^{-\nu};$$

and we conclude as we did for Ψ. □

Proposition 7.2. *Assume* (1.4), (7.1), (7.2), *and* (7.3). *Let* Ψ *be given by* (7.7) *and let* $u^+ = c\Psi$ *where the constant* c *is small enough so that* u^+ *satisfies the hypotheses of part* (iii) *of Theorem* 5.1 *with* $\mu = \nu$ *and let* $u \in X_\alpha \cap \mathcal{X}_{\alpha,\nu} \cap C([0,\infty), H^1(\mathbb{R}^N))$ *be the resulting solution of* (1.1). *It follows that*

$$t^\nu \|u(t) - \mathcal{T}(t)u^+\|_{L^{\alpha+2}} + t^\nu \|\mathcal{T}(-t)u(t) - u^+\|_{L^{\alpha+2}} \leq Ct^{-\alpha(\nu-\beta)}, \tag{7.11}$$

and

$$t^\nu \|u(t) - c\mathcal{T}(t)\psi\|_{L^{\alpha+2}} \leq C\max\{t^{-\alpha(\nu-\beta)}, t^{-\frac{N\alpha}{2(\alpha+2)}+\nu}\} \tag{7.12}$$

for all $t > 0$. *In particular,*

$$0 < \liminf_{t\to\infty} t^\nu \|u(t)\|_{L^{\alpha+2}} \leq \limsup_{t\to\infty} t^\nu \|u(t)\|_{L^{\alpha+2}} < \infty. \tag{7.13}$$

In addition,

$$\|u(t) - \mathcal{T}(t)u^+\|_{H^1} \leq Ct^{-\alpha(\nu-\beta)}. \tag{7.14}$$

The same is true if $u^+ = c\Theta$ *where* Θ *is given by* (7.8) *and* c *is sufficiently small.*

Proof. Recall that $u \in C([0,\infty), H^1(\mathbb{R}^N))$ by Proposition 5.4. The first estimate is a consequence Proposition 3.7 (i) and (iii). Furthermore, since $c\psi = u^+ + c\eta\psi$, it follows that

$$\begin{aligned} t^\nu \|u(t) - \mathcal{T}(t)c\psi\|_{L^{\alpha+2}} &\leq t^\nu \|u(t) - \mathcal{T}(t)u^+\|_{L^{\alpha+2}} + t^\nu \|\mathcal{T}(t)c\eta\psi\|_{L^{\alpha+2}} \\ &\leq Ct^{-\alpha(\nu-\beta)} \|u\|_{\mathcal{X}_{\alpha,\nu}} + Ct^{-\frac{N\alpha}{2(\alpha+2)}+\nu} \|\eta\psi\|_{L^{\frac{\alpha+2}{\alpha+1}}}. \end{aligned}$$

This proves the second estimate, which immediately implies the third. The last one follows from Proposition 3.10 (iv). If $u^+ = c\Theta$, the same argument works using $c\psi = u^+ + c(\psi - \Theta)$. $\square$

REMARK 7.3. (7.13) gives the precise decay rate of $\|u(t)\|_{L^{\alpha+2}}$ as $t \to \infty$. Notice that for different choices of q, we obtain H^1 solutions of (1.1) with different decay rates. By (7.1), (7.2), and (7.3), the decay rates which can be realized by Proposition 7.2 are

$$\beta < \nu < \frac{N\alpha}{2(\alpha+2)} \quad \text{if} \quad \alpha_0 < \alpha \le \frac{4}{N} \tag{7.15}$$

and

$$\frac{N\alpha}{4(\alpha+2)} < \nu < \frac{N\alpha}{2(\alpha+2)} \quad \text{if} \quad \frac{4}{N} \le \alpha < \frac{4}{N-2}. \tag{7.16}$$

Note that $\frac{N\alpha\alpha}{4(\alpha+2)} < \beta$ precisely when $\alpha < 4/N$. We can make the following additional comments on the optimality of (7.15) and (7.16).

(i) The right-hand limit $\nu = \frac{N\alpha}{2(\alpha+2)}$ in (7.15) and (7.16) can be achieved for solutions with initial values $\varphi \in H^1(\mathbb{R}^N) \cap L^2(\mathbb{R}^N, |x|^2 dx)$ (see [16] and [3]). On the other hand, Corollary 3.9 shows that it cannot be improved.

(ii) If $\alpha > 4/N$, then Remark 3.12 shows that the lower bound in (7.16) cannot even be attained.

(iii) We conjecture that the lower bound in (7.15) is not sharp. Indeed, if $\alpha_0 < \alpha < 4/N$ and $N/2 < \operatorname{Re} q < 2/\alpha$, then Ψ defined by (7.7) is in $H^1(\mathbb{R}^N)$ but not in W_α. Thus the resulting (global) solution is not in X_α. On the other hand, by Corollary 7.4.6 in [1], at least in the radially symmetric case and when $\gamma > 0$, $\|u(t)\|_{L^{\alpha+2}} \to 0$ as $t \to \infty$.

REMARK 7.4. Under the hypotheses of Proposition 7.2, (1.1) has a large class of solutions, corresponding to different choices of the cut-off function η, all of which are asymptotic as $t \to \infty$ to the same self-similar solution of the linear Schrödinger equation (whose scaling properties are different from those of (1.1)). See Theorems 1.4 and 1.5 in [17] for similar results for the "two-power" nonlinear Schrödinger equation.

REMARK 7.5. If, in addition to the hypotheses of Proposition 7.2, we assume $\alpha > 4/N$, then the property that $u(t) - \mathcal{T}(t)u^+ \to 0$ in $H^1(\mathbb{R}^N)$ is a well-known result of the "classical" H^1 theory of Ginibre and Velo [6]. The precise decay estimates (7.11), (7.13), and (7.14) are new as far as we are aware. In other words, new information about H^1 solutions of (1.1) has been obtained by studying X_α solutions.

Applying Proposition 5.5 instead of Theorem 5.1, we obtain the following analogue of Proposition 7.2, where we do not need the smallness condition, but instead we assume $\gamma > 0$ or $\alpha < 4/N$.

PROPOSITION 7.6. *Assume* (1.4), (7.1), (7.2), *and* (7.3). *Suppose in addition* $\gamma > 0$ *or* $\alpha < 4/N$. *Let* Ψ *be given by* (7.7) *and let* $u^+ = c\Psi$ *for some* $c \in \mathbb{C}$. *Let* $u \in C([0,\infty), H^1(\mathbb{R}^N)) \cap \mathcal{X}_{\alpha,\nu}$ *be the resulting solution of* (1.1), *as given by Proposition* 5.5. *It follows that* (7.11), (7.12), (7.13), *and* (7.14) *hold. The same is true if* $u^+ = c\Theta$, *where* Θ *is given by* (7.8).

The next proposition is the analogue of Proposition 7.2 for initial values instead of scattering states. Its proof uses Theorem 2.1 in [4] to guarantee the existence of global solutions u of (1.1) in $X_\alpha \cap \mathcal{X}_{\alpha,\nu}$. This theorem requires the additional condition $\alpha\beta + \nu < 1$ in order to conclude $u \in \mathcal{X}_{\alpha,\nu}$. In the present situation, this condition is automatic, since by (7.6) and (2.7) $\nu < \frac{N\alpha}{2(\alpha+2)} = 1 - \alpha\beta$.

Proposition 7.7. *Assume* (1.4), (7.1), (7.2), *and* (7.3). *Let Ψ be given by* (7.7) *and $\varphi = c\Psi$ where the constant c is small enough so that φ satisfies the hypotheses of Theorem* 2.1 *in* [4] *with $\delta = \nu - \beta$. Let $u \in X_\alpha \cap \mathcal{X}_{\alpha,\nu}$ be the resulting solution of* (1.1), *which is also a "classical" H^1 solution by Proposition* 2.3 *in* [4]. *It follows that*

$$t^\nu \|u(t) - c\mathcal{T}(t)\psi\|_{L^{\alpha+2}} = O(t^{-\varepsilon}) \tag{7.17}$$

as $t \to \infty$ for some $\varepsilon > 0$. In particular, u satisfies (7.13). *Moreover, u satisfies* (7.11) *and* (7.14), *where u^+ is the scattering state of u. These results are true if $\varphi = c\Theta$ where Θ is given by* (7.8) *and c is sufficiently small.*

Proof. If $(\alpha+1)\nu < 1$, the proof of (7.17) is similar to the proof of (7.12) using part (ii) of Proposition 3.7. If $(\alpha+1)\nu \geq 1$, fix $\nu' \in (\beta, \nu)$ such that $(\alpha+1)\nu' < 1$. By using part (ii) of Proposition 3.7 again, we obtain

$$\begin{aligned} t^\nu \|u(t) - \mathcal{T}(t)c\psi\|_{L^{\alpha+2}} &\leq t^\nu \|u(t) - \mathcal{T}(t)\varphi\|_{L^{\alpha+2}} + t^\nu \|\mathcal{T}(t)c\eta\psi\|_{L^{\alpha+2}} \\ &\leq Ct^{\nu-\nu'-\alpha(\nu'-\beta)} \|u\|^{\alpha+1}_{\mathcal{X}_{\alpha,\nu'}} + Ct^{-\frac{N\alpha}{2(\alpha+2)}+\nu} \|\eta\psi\|_{L^{\frac{\alpha+2}{\alpha+1}}}. \end{aligned}$$

It follows from (7.6) that $-\frac{N\alpha}{2(\alpha+2)} + \nu < 0$ and that

$$\nu - \nu' - \alpha(\nu' - \beta) = \nu - (\alpha+1)\nu' + 1 - \frac{N\alpha}{2(\alpha+2)} < 0$$

if $(\alpha+1)\nu'$ is sufficiently close to 1. This proves (7.17). (7.11) follows from Proposition 3.7 (i) and (iii), and (7.14) follows from Proposition 3.10 (iv). This completes the proof if $\varphi = c\Psi$. The proof for $\varphi = c\Theta$ is analogous. □

8. Appendix.

In this appendix, we give some properties of the linear Schrödinger equation.

Proposition 8.1. *If $\varphi \in \mathcal{S}'(\mathbb{R}^N)$, $\varphi \neq 0$, then*

$$\liminf_{t\to\infty} t^{\frac{N\alpha}{2(\alpha+2)}} \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} > 0$$

for all $\alpha \geq 0$. Here, we use the convention that $\|\psi\|_{L^{\alpha+2}} = +\infty$ if $\psi \in \mathcal{S}'(\mathbb{R}^N)$, $\psi \notin L^{\alpha+2}(\mathbb{R}^N)$.

Proof. We claim that

$$\mathcal{T}(-t)\mathcal{F}\psi = i^{\frac{N}{2}} D_{\frac{1}{4\pi t}} M_{-16\pi^2 t} \mathcal{T}\Big(\frac{1}{16\pi^2 t}\Big)\psi \tag{8.1}$$

for all $t > 0$ and all $\psi \in \mathcal{S}'(\mathbb{R}^N)$, where the dilation D_ℓ and the multiplier M_μ are defined by

$$(D_\ell\theta)(x) = \ell^{\frac{N}{2}}\theta(\ell x), \quad (M_\mu\theta)(x) = e^{i\frac{\mu|x|^2}{4}}\theta(x)$$

for all $\theta \in \mathcal{S}(\mathbb{R}^N)$ and by duality for all $\theta \in \mathcal{S}'(\mathbb{R}^N)$. Indeed, by density it suffices to establish (8.1) for $\psi \in \mathcal{S}(\mathbb{R}^N)$. Since $\mathcal{T}(t) = \mathcal{F}M_{-16\pi^2 t}\mathcal{F}^{-1}$ (take the inverse Fourier transform of the Schrödinger equation satisfied by $\mathcal{T}(t)\psi$), it follows that $\mathcal{T}(-t)\mathcal{F} = \mathcal{F}M_{16\pi^2 t}$. Next, we use the following formula (see formula (3.5) in [2]) which can be verified by a direct calculation with the kernel of $\mathcal{T}(t)$,

$$\mathcal{F}M_{16\pi^2 t} = i^{\frac{N}{2}} D_{\frac{1}{4\pi t}} M_{-16\pi^2 t} \mathcal{T}\Big(\frac{1}{16\pi^2 t}\Big).$$

(8.1) follows. We now deduce from (8.1) that

$$\|\mathcal{T}(-t)\mathcal{F}\psi\|_{L^{\alpha+2}} = (4\pi t)^{-\frac{N\alpha}{2(\alpha+2)}} \Big\|\mathcal{T}\Big(\frac{1}{16\pi^2 t}\Big)\psi\Big\|_{L^{\alpha+2}} \tag{8.2}$$

for all $\psi \in \mathcal{S}'(\mathbb{R}^N)$.

Finally, suppose by contradiction that there exist $\alpha \geq 0$ and a sequence $\tau_n \to \infty$ such that

$$\tau_n^{\frac{N\alpha}{2(\alpha+2)}} \|\mathcal{T}(\tau_n)\varphi\|_{L^{\alpha+2}} \underset{n\to\infty}{\longrightarrow} 0.$$

Applying (8.2) with $t = (16\pi^2\tau_n)^{-1}$, we deduce that

$$\Big\|\mathcal{T}\Big(-\frac{1}{16\pi^2\tau_n}\Big)\mathcal{F}\varphi\Big\|_{L^{\alpha+2}} \underset{n\to\infty}{\longrightarrow} 0.$$

Since $\tau_n \to \infty$, this implies that $\mathcal{F}\varphi = 0$ in $\mathcal{S}'(\mathbb{R}^N)$, thus $\varphi = 0$. □

Remark 8.2. For every $\alpha > 0$, the following properties hold:

(i) Given $0 < \nu < \frac{N\alpha}{2(\alpha+2)}$, there exists $\varphi \in \mathcal{S}'(\mathbb{R}^N)$ such that $t^\nu \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} = 1$ for all $t > 0$. Indeed, let $\varphi(x) = c|x|^{-2\nu - \frac{N}{\alpha+2}}$ and apply Proposition 3.7 in [4].

(ii) Given $\frac{N\alpha}{4(\alpha+2)} < \nu < \frac{N\alpha}{2(\alpha+2)}$, there exist $0 < a < b < \infty$ and $\varphi \in H^\infty(\mathbb{R}^N)$ such that $a \leq t^\nu \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} \leq b$ for all $t \geq 1$. Indeed, let $\varphi(x) = c|x|^{-2\nu-\frac{N}{\alpha+2}}\theta(x)$ with $\theta \in C^\infty(\mathbb{R}^N)$, $\theta(x) = 0$ for $|x| \leq 1$, and $\theta(x) = 1$ for $|x| \geq 2$.

Remark 8.3. Here are a few observations on the structure of the Banach space $\mathcal{W}_{\alpha,\mu}$.

(i) $(\mathcal{T}(t))_{t\geq 0}$ is a semigroup of contractions on $\mathcal{W}_{\alpha,\mu}$, but if $0 < \mu < \frac{N\alpha}{2(\alpha+2)}$, then $(\mathcal{T}(t))_{t\geq 0}$ is **not** continuous. More precisely, if $\varphi \in \mathcal{W}_{\alpha,\mu}$, then one verifies easily that

$$\|\mathcal{T}(\tau)\varphi - \varphi\|_{\mathcal{W}_{\alpha,\mu}} \geq \liminf_{t\downarrow 0} t^\mu \|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}}$$

for all $\tau > 0$. In particular, if $\varphi(x) = |x|^{-2\mu-\frac{N}{\alpha+2}}$, then $\|\mathcal{T}(\tau)\varphi - \varphi\|_{\mathcal{W}_{\alpha,\mu}} \geq \|\varphi\|_{\mathcal{W}_{\alpha,\mu}} > 0$ for all $\tau > 0$ by Remark 8.2 (i).

(ii) If $0 < \mu < \frac{N\alpha}{2(\alpha+2)}$, then $\mathcal{S}(\mathbb{R}^N)$ is **not** dense in $\mathcal{W}_{\alpha,\mu}$. More precisely, if $\varphi(x) = |x|^{-2\mu-\frac{N}{\alpha+2}}$ and $\theta \in \mathcal{S}(\mathbb{R}^N)$, then $\|\varphi - \theta\|_{\mathcal{W}_{\alpha,\mu}} \geq \|\varphi\|_{\mathcal{W}_{\alpha,\mu}} > 0$. Indeed, by Remark 8.2 (i),

$$t^\mu\|\mathcal{T}(t)(\varphi-\theta)\|_{L^{\alpha+2}} \geq t^\mu\|\mathcal{T}(t)\varphi\|_{L^{\alpha+2}} - t^\mu\|\mathcal{T}(t)\theta\|_{L^{\alpha+2}} = \|\varphi\|_{\mathcal{W}_{\alpha,\mu}} - t^\mu\|\mathcal{T}(t)\theta\|_{L^{\alpha+2}},$$

and the conclusion follows by letting $t \downarrow 0$. We do not know what is the closure of $\mathcal{S}(\mathbb{R}^N)$ in $\mathcal{W}_{\alpha,\mu}$.

(iii) If $0 < \mu < \frac{N\alpha}{2(\alpha+2)}$, then $\mathcal{W}_{\alpha,\mu}$ is **not** separable. To see this, let $\varphi(x) = |x|^{-2\mu-\frac{N}{\alpha+2}}$ and, given $v \in \mathbb{R}^N$, set $\varphi_v(x) = e^{i\frac{v\cdot x}{2}}\varphi(x)$. If $v \neq w$, we claim that $\|\varphi_v - \varphi_w\|_{\mathcal{W}_{\alpha,\mu}} \geq \|\varphi\|_{\mathcal{W}_{\alpha,\mu}}$, which implies the result. Indeed, by Galilean invariance, $(\mathcal{T}(t)\varphi_v)(x) = e^{-it\frac{|v|^2}{4}}e^{i\frac{v\cdot x}{2}}(\mathcal{T}(t)\varphi)(x - tv)$. Thus $\varphi_v \in \mathcal{W}_{\alpha,\mu}$ and $\|\varphi_v\|_{\mathcal{W}_{\alpha,\mu}} = \|\varphi\|_{\mathcal{W}_{\alpha,\mu}}$. Moreover, if f is the profile associated with φ, then

$$\mathcal{T}(t)(\varphi_v - \varphi_w)(x) = t^{-\mu-\frac{N}{2(\alpha+2)}}\left[e^{-it\frac{|v|^2}{4}}e^{i\frac{v\cdot x}{2}}f\Big(\frac{x-tv}{\sqrt{t}}\Big) - e^{-it\frac{|w|^2}{4}}e^{i\frac{w\cdot x}{2}}f\Big(\frac{x-tw}{\sqrt{t}}\Big)\right].$$

In particular,

$$t^{\mu}\|\mathcal{T}(t)(\varphi_v - \varphi_w)\|_{L^{\alpha+2}(\{|x-tv|<1\})} \geq \|f\|_{L^{\alpha+2}(\{|x|<\sqrt{t}\})} - \|f\|_{L^{\alpha+2}(\{|x|>\sqrt{t}|v-w|-1\})}.$$

Letting $t \to \infty$, we see that

$$\liminf_{t\to\infty} t^{\mu}\|\mathcal{T}(t)(\varphi_v - \varphi_w)\|_{L^{\alpha+2}(\{|x-tv|<1\})} \geq \|f\|_{L^{\alpha+2}},$$

and the claim follows.

REFERENCES

[1] T. Cazenave, *An introduction to nonlinear Schrödinger equations*, 3rd ed., Text. Métod. Mat. 26, Instituto de Matemática, Universidade Federal de Rio de Janeiro, Rio de Janeiro, 1996.

[2] T. Cazenave and F. B. Weissler, *The structure of solutions to the pseudo-conformally invariant nonlinear Schrödinger equation*, Proc. Royal Soc. Edinburgh Sect. A, 117 (1991), pp. 251–273.

[3] T. Cazenave and F. B. Weissler, *Rapidly decaying solutions of the nonlinear Schrödinger equation*, Comm. Math. Phys., 147 (1992), pp. 75–100.

[4] T. Cazenave and F. B. Weissler, *Asymptotically self-similar global solutions of the nonlinear Schrödinger and heat equations*, Math. Z., 228 (1998), pp. 83–120.

[5] T. Cazenave and F. B. Weissler, *More self-similar solutions of the nonlinear Schrödinger equation*, Nonlinear Differential Equations Appl., 5 (1998), pp. 355–365.

[6] J. Ginibre and G. Velo, *Scattering theory in the energy space for a class of nonlinear Schrödinger equations*, J. Math. Pures Appl., 64 (1985), pp. 363–401.

[7] R. Johnson and X. Pan, *On an elliptic equation related to the blow-up phenomenon in the nonlinear Schrödinger equation*, Proc. Royal Soc. Edinburgh Sect. A, 123 (1993), pp. 763–782.

[8] T. Kato, *Nonlinear Schrödinger equations*, in Schrödinger Operators, Lecture Notes in Phys. 345, Springer, New York, 1989, pp. 218–263.

[9] O. Kavian and F. B. Weissler, *Self-similar solutions of the pseudo-conformally invariant nonlinear Schrödinger equation*, Michigan Math. J., 41 (1992), pp. 151–173.

[10] N. Kopell and M. Landman, *Spatial structure of the focusing singularity of the nonlinear Schrödinger equation: A geometrical analysis*, SIAM J. Appl. Math., 55 (1995), pp. 1297–1323.

[11] Tzong-Yow Lee and Wei-Ming Ni, *Global existence, large time behavior and life span of solutions of a semilinear parabolic Cauchy problem*, Trans. Amer. Math. Soc., 333 (1992), pp. 365–378.

[12] F. Oru, *Rôle des oscillations dans quelques problèmes d'analyse non-linéaire*, Ph.D. thesis, Ecole Normale Supérieure de Cachan, Cachan, France, 1998.

[13] F. Ribaud and A. Youssfi, *Regular and self-similar solutions of nonlinear Schrödinger equations*, J. Math. Pures Appl., 77 (1998), pp. 1065–1079.

[14] W. Rudin, *Functional Analysis*, McGraw–Hill, New York, 1991.

[15] E. M. Stein, *Harmonic Analysis: Real Variable Methods, Orthogonality and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[16] Y. Tsutsumi, *Scattering problem for nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Phys. Théor., 43 (1985), pp. 321–347.

[17] F. B. Weissler, *Asymptotically Self-Similar Solutions of the Two-Power Nonlinear Schrödinger Equation*, manuscript.

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 651–677

AN ELLIPTIC REGULARITY RESULT FOR A COMPOSITE MEDIUM WITH "TOUCHING" FIBERS OF CIRCULAR CROSS-SECTION[*]

ERIC BONNETIER[†] AND MICHAEL VOGELIUS[‡]

Abstract. In this paper we consider the elliptic equation $\nabla \cdot a\nabla u = 0$ in a two dimensional domain Ω, which contains a finite number of circular inhomogeneities (cross-sections of fibers). The coefficient, a, takes two constant values, one in all the inhomogeneities and one in the part of Ω which lies outside the inhomogeneities. A number of the inhomogeneities may possibly touch, but in spite of this we prove that any variational solution u (with sufficiently smooth boundary data) is in $W^{1,\infty}$. For this very interesting, particular type of coefficient, our result improves a classical regularity result due to DeGiorgi and Nash, which asserts that the solution is in the Hölder class C^γ for some positive exponent γ.

Key words. elliptic regularity, composite materials

AMS subject classifications. 35J25, 73C40

PII. S0036141098333980

1. Introduction.

Consider a domain $\Omega \subset \mathbb{R}^2$, representing the cross-section of a three dimensional body. We suppose the three dimensional body is occupied by a fiber-reinforced composite and we suppose the cross-section is taken perpendicular to the finitely many (identical) cylindrical fibers. Frequently in composites, the fibers are very closely spaced and may even touch. We suppose the strength of the fibers is different from that of the material between the fibers (the so-called matrix). When we talk about strength this could, for instance, refer to the shear modulus (for a problem of antiplane shear) or the conductivity (for a problem of heat or electric conduction). In all three cases the corresponding scalar variable, u, (the out of plane displacement, the temperature or the voltage potential) satisfies the elliptic equation

$$\nabla \cdot a\nabla u = 0 \quad \text{in } \Omega, \tag{1}$$

with, for instance, a given Dirichlet boundary condition

$$u = \phi \quad \text{on } \partial\Omega. \tag{2}$$

The coefficient $0 < a < \infty$ takes two constant values, one in the fibers and one in the matrix. The aim of this paper is to study the behavior of u and, in particular, its gradient near points where the fibers touch. Because the cross-section is perpendicular to the fibers, these appear as disks of identical radii. We may without loss of generality restrict attention to a situation with only two touching disks (fibers); for simplicity we take these to lie strictly inside Ω. We may also rescale the strength, a, so that

$$\begin{aligned} &a(x) = 1 \quad \text{for } x \text{ outside the two disks,} \\ &a(x) = a_0 \quad \text{for } x \text{ inside the two disks.} \end{aligned} \tag{3}$$

[*] Received by the editors February 12, 1998; accepted for publication (in revised form) January 7, 1999; published electronically March 6, 2000.
http://www.siam.org/journals/sima/31-3/33398.html

[†] CNRS, CMAP, Ecole Polytechnique, 91128 Palaiseau, France (eopus@cmapx.polytechnique.fr).

[‡] Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (vogelius@math.rutgers.edu). This author was supported by National Science Foundation grant DMS-9704575.

In the context of antiplane shear it is probably physically most relevant to think of a_0 as being larger than 1—after all, the fibers are there for reinforcement. However, in the context of heat or electrical conduction there are no physical reasons why we might not have $a_0 < 1$ as well. We may, without loss of generality, suppose that the point where the two disks touch is located at the origin. We may also suppose that the domain Ω is of class C^∞ and symmetric in the x_1-axis, and that the boundary data ϕ is in $C^\infty(\partial\Omega)$. If not, we can simply take such a smooth domain inside Ω, containing the two fiber cross-sections and rely on elliptic regularity to get that the (new) Dirichlet data is C^∞. The geometric situation is illustrated in Figure 1. For simplicity of illustration, we have drawn $\partial\Omega$ in the shape of a circle—a convention we shall follow throughout this paper. The solution u is defined variationally by the requirements that u be in $H^1(\Omega)$ with $u|_{\partial\Omega} = \phi$ and

$$\int_\Omega a(x)\nabla u \nabla v \, dx = 0 \quad \forall v \in H_0^1(\Omega).$$

In two dimensions (as we are here) Sobolev's imbedding theorem states that elements of $H^s(\Omega)$, $1 < s$ are automatically continuous. This is not true for all elements of $H^1(\Omega)$. However, a regularity result of DeGiorgi and Nash (cf. [3], [10], or [4]) asserts that any H^1 solution to a divergence form, scalar elliptic equation with bounded measurable coefficients such as (1), is indeed Hölder continuous inside Ω. Near $\partial\Omega$, the coefficent a is constant, and the boundary, as well as the boundary data ϕ, are C^∞, so standard elliptic boundary regularity results immediately imply that u is C^∞ there. Indeed, away from the origin (where the two disks touch) standard elliptic regularity results (for operators with constant, or piecewise constant coefficients) very easily imply that the gradient of u is bounded. The origin, however, presents a serious problem. Neither standard elliptic regularity results nor the DeGiorgi–Nash result assert anything about the boundedness of the gradient. Such boundedness is guaranteed by the main result of this paper.

Theorem. *The solution u is in $W^{1,\infty}(\Omega)$ for any fixed $0 < a_0 < \infty$.*

Since the gradient of the solution u is generically discontinuous at the interfaces between the fibers and the matrix, this theorem is optimal in terms of global regularity.

We have assumed that the circular fiber cross-sections have the same radius. It is worthwhile to point out that this assumption is for convenience only. A configuration with two touching disks of different radii (say, r_1 and r_2) may quite easily be mapped conformally to a configuration consisting of two identical touching disks: pick the x_1-axis to be the common tangent for the two disks (they touch at the origin) and let $z = \Phi(x)$ denote the conformal mapping $\Phi(x_1, x_2) = (x_1/(x_1^2 + x_2^2), x_2/(x_1^2 + x_2^2))$. Let T_c denote the vertical translation $T(z_1, z_2) = (z_1, z_2 + c)$. With an appropriate choice of c $(= \pm(r_2 - r_1)/4r_1r_2)$ the conformal mapping $\Psi = \Phi^{-1} \circ T_c \circ \Phi$ maps the configuration with the two different disks to a configuration with two identical touching disks (of radius $2r_1r_2/(r_1 + r_2)$). This mapping is furthermore smooth at the origin. The validity of the above theorem for two identical disks now immediately implies its validity for different disks as well: the "push-forward" $w = u \circ \Psi^{-1}$ of the function u is in $W^{1,\infty}$ near the origin (since it pertains to a configuration of two identical disks) and due to the regularity of Ψ the same can therefore be said about u. This "push-forward" technique works for any configuration which is conformally equivalent to two touching disks. It should be extremely interesting to study the regularity issue for configurations that are not conformally equivalent to two touching disks.

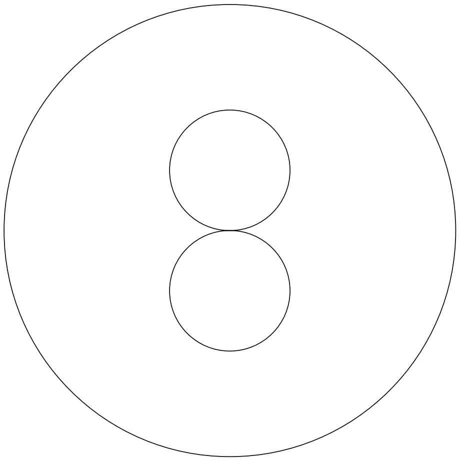


FIG. 1. *Two touching fibers.*

In the context of antiplane shear, $a\nabla u$ represents the stresses (internal forces). Most linear fracture models suppose that fracturing will occur at points with extreme stress concentrations. The fact that the (shear) stresses remain bounded, even near points where the fibers touch, strongly indicates that separation between circular fibers and the matrix is not a likely mechanism for the onset of fracture. The result proven in this paper only applies to a scalar equation; it would be of utmost interest to extend this to the full system of linear elasticity.

To indicate that the behavior of u near the origin is not entirely obvious (and not always the same) let us change to the (conformally) different situation shown in Figure 2: two conical shapes symmetrically touching, with $a = a_0$ inside the two shapes, $a = 1$ outside. For a fixed a_0, the solution is in C^γ with γ uniformly bounded away from zero (independently of the size of the interior angle of the conical shapes). This is consistent with the result of DeGiorgi and Nash, which asserts that u is of a minimal Hölder class that only depends on the aspect ratio of the coefficients. The actual Hölder exponent γ is "generically" smallest when the conical shapes have interior angles $\alpha = \pi/2$. Corresponding to this exponent there is locally near the origin an H^1 solution to (1), which in polar coordinates has the form

$$r^\gamma(A\cos\gamma\theta + B\sin\gamma\theta),$$

with a different pair of constants (A, B) in each of the four sectors. γ is the smallest exponent for which a solution of this form exists in H^1. γ is very easily characterized as the smallest positive solution to a certain determinant identity (expressing that the system of linear equations for the coefficient pairs (A, B) has a nontrivial solution). An analysis very similar to that found in [5] would show that near the origin any solution has, generically, the same behavior as this special solution. Figure 3 shows the "generic γ" ($\alpha = \pi/2$) as a function of $a_0 \neq 1$ for the two possible symmetries of the solution u: one where u is odd with respect to the x_1 axis, and one where u is even with respect to the x_1 axis. It is clear from the graph that a general solution u, which contains elements of both symmetries, is never in $W^{1,\infty}$ (except when $a_0 = 1$); indeed, depending on a_0, its regularity may not be better than C^γ for any arbitrarily small positive γ.

The fact that u is not in $W^{1,\infty}(\Omega)$ in this situation can be explained in terms of

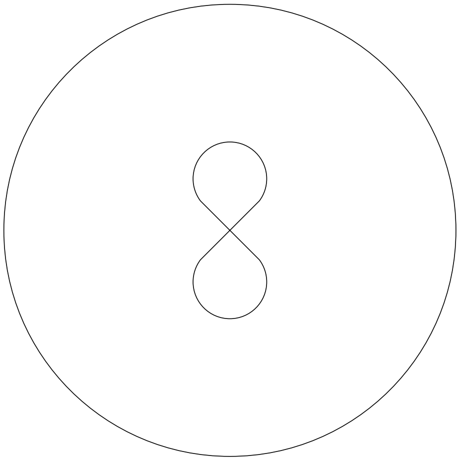


FIG. 2. *Two touching conical shapes.*

a "corner effect." Consider the geometric situation where the two convex shapes are $\epsilon > 0$ apart vertically. The solution will then have two singularities which arise due to the corners in the interfaces. Each corner has an angle of $\pi/2$. Figure 4 shows the generic Hölder coefficient for any solution corresponding to an interface with angle $\pi/2$ (and conductivity ratio $1 : a_0$). Figure 4 corresponds to the lower half of Figure 3; if we had considered solutions with a special symmetry, the curve on the right side (of $a_0 = 1$) would continue smoothly into the left quadrant, and vice versa. It is clear that, for $\epsilon > 0$, a general solution is never in $W^{1,\infty}(\Omega)$ (except when $a_0 = 1$), and it is quite natural to expect that this does not in any way improve in the limit $\epsilon \to 0$, when the two shapes touch.

But the "corner effect" is not the whole story in Figure 3. The regularity situation has clearly become significantly worse when compared to that in Figure 4. For fixed $\epsilon > 0$, Figure 4 clearly indicates that any solution will at least be in $C^{2/3}(\Omega)$, independently of a_0. However, when the shapes touch (for $\epsilon = 0$) Figure 3 clearly indicates that the regularity may not be better than C^γ for any arbitrarily small positive γ, depending on the size of a_0. The touching may thus induce a loss of almost a factor of 2/3 in terms of differentiability.

We now return to the case of (identical) circular fibers. The extreme situations that formally correspond to $a_0 = 0$ and $a_0 = \infty$ are somewhat particular. The corresponding solutions are now (essentially) only defined in $\Omega \setminus \{\text{the fibers}\}$. If the boundary point where the two fibers touch (the origin) is thought of as two different boundary points of $\Omega \setminus \{\text{the fibers}\}$, then these solutions are always C^∞ in the interior and up to the boundary of $\Omega \setminus \{\text{the fibers}\}$. With this convention we also have that all the derivatives of order ≥ 1 vanish at the origin. In the case $a_0 = 0$, we generically have that the solution u^0 is multivalued at the origin, i.e., it has a different limit when approaching the origin from the cusp on the left than when approaching the origin from the cusp on the right. For $a_0 = \infty$, the solution u^∞ is always single valued at the origin. We refer the reader to the appendix, where these issues are discussed in more detail.

An interesting project would be to consider the case where the fibers are very close but not touching. Say, the circular (unit size) cross-sections are ϵ apart vertically. For the case of $a_0 = 0$ (as well as $a_0 = \infty$) the discontinuity mentioned above gives the

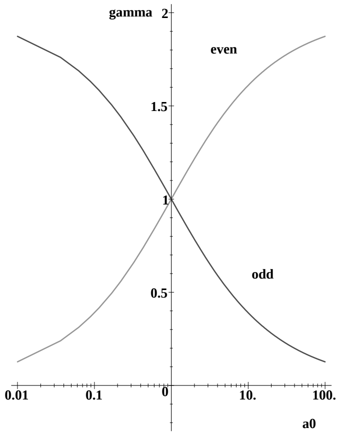


FIG. 3. *The generic γ as a function of a_0 for symmetrically touching conical shapes of interior angle $\pi/2$.*

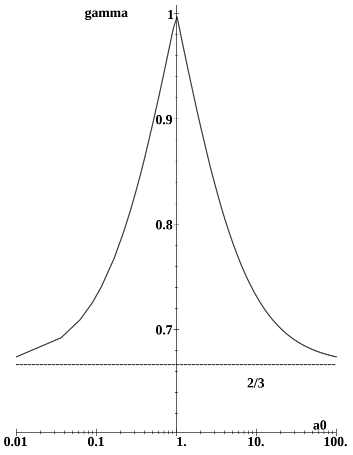


FIG. 4. *The generic γ as a function of a_0 for a single conical shape of interior angle $\pi/2$.*

existence of solutions, the gradients of which become unbounded at the origin as the distance ϵ approaches zero. This phenomenon has been noted and studied in detail by several authors (cf. [2], [9], and [6]). We again refer to the appendix, where some of this work is discussed in a little more detail. For fixed $0 < a_0 < \infty$, we conjecture that the gradient near the origin stays bounded independently of ϵ. We base this conjecture, among other things, on some very accurate calculations communicated to us by Börje Andersson of the Aeronautical Research Institute of Sweden [1]. For circular cross-sections and special boundary conditions, such as those corresponding to the solutions considered in [2] and [9], it is not unlikely that the mapping technique used there would make it possible to establish this uniform boundedness. However, for general boundary conditions or for more general (smooth) fiber cross-sections, entirely different techniques will be required to prove the conjecture.[1]

2. A reduced problem. Since Ω is symmetric in the x_1-axis, the boundary condition $u|_{\partial\Omega} = \phi(x)$ may also be written as $u|_{\partial\Omega} = \frac{1}{2}(\phi(x) - \phi(\overline{x})) + \frac{1}{2}(\phi(x) + \phi(\overline{x}))$, where $\overline{x} = (x_1, -x_2)$ denotes the reflection of the point $x = (x_1, x_2)$. As a consequence, we may separate our boundary value problem into one of two cases: (1) the solution u is odd in the x_1-axis; (2) the solution u is even in the x_1-axis.

In case the boundary data (and thus the solution, u) is even in the x_1-axis, consider the a-harmonic conjugate to u. This function, v, is related to u by

$$a\nabla u = \nabla v^{\perp} = \begin{pmatrix} -\frac{\partial v}{\partial x_2} \\ \frac{\partial v}{\partial x_1} \end{pmatrix}, \tag{4}$$

and it solves

$$\nabla \cdot a^{-1}(x)\nabla v = 0 \quad \text{in } \Omega.$$

From (4) it follows immediately that on $\partial\Omega$

$$\begin{aligned} a^{-1}\nabla v \cdot n &= -a^{-1}\nabla v \cdot \tau^{\perp} \\ &= a^{-1}\nabla v^{\perp} \cdot \tau \\ &= \frac{\partial u}{\partial \tau}, \end{aligned}$$

where τ denotes the counterclockwise tangent. Since u is even it now follows that v (normalized by $\int_\Omega v\,dx = 0$) is odd in the x_1-axis.

In the rest of this paper we shall concentrate on the case where u is odd in the x_1-axis and prove that the gradient of u is bounded. The exact same argument that we present could of course be applied to the (odd) function v (the only difference being that a_0, the conductivity inside the two fibers, gets replaced by a_0^{-1}), thus proving that the gradient of v stays bounded in Ω. Because of the relationship (4) this immediately implies that ∇u is also bounded in Ω in the case where u is even in the x_1-axis. By means of the splitting introduced at the beginning of this section, this now verifies the boundedness of ∇u in the general case (without any symmetry assumptions).

Under the assumption that u is odd in the x_1-axis it indeed suffices to consider the boundary value problem in the half-domain, $\Omega_+ = \{(x_1, x_2) \in \Omega : 0 < x_2\}$, with the additional boundary condition $u = 0$ at $x_2 = 0$ (cf. Figure 5). For simplicity we shall from now on assume that the fiber has radius 1, so that its boundary is given by the equation $x_1^2 + (x_2^2 - 1)^2 = 1$.

[1] Quite recently Y.-Y. Li and M. Vogelius have developed such techniques and among other results established that the conjecture is indeed true; see [8].

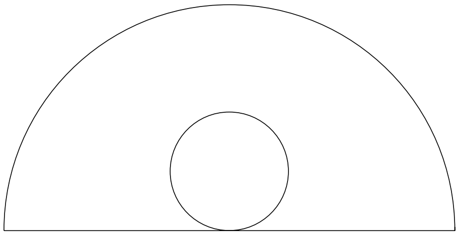


FIG. 5. *The reduced geometric setup.*

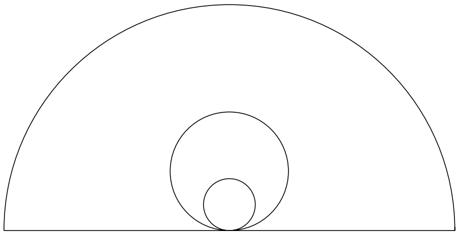


FIG. 6. *The geometric situation with the "auxiliary" boundary.*

3. An auxiliary boundary value problem. Let D_ϵ denote the disk $D_\epsilon = \{(x_1, x_2) : x_1^2 + (x_2 - \epsilon)^2 \le \epsilon^2\}$, centered at $(0, \epsilon)$, with radius ϵ. For small, positive ϵ we introduce an auxiliary function u_ϵ as the solution to the boundary value problem

$$\begin{aligned} \nabla \cdot a(x)\nabla u_\epsilon = 0 &\quad \text{in } \Omega_+ \setminus D_\epsilon, \\ u_\epsilon = \phi &\quad \text{on } \partial\Omega_+ \setminus \{x_2 = 0\}, \\ u_\epsilon = 0 &\quad \text{on } \{x_2 = 0\} \text{ and on } \partial D_\epsilon. \end{aligned}$$

For a geometric illustration, see Figure 6.
The conductivity, a, is as given before. The solution, u, whose gradient we are trying to bound, solves the corresponding "limiting" boundary value problem

$$\begin{aligned} \nabla \cdot a(x)\nabla u = 0 &\quad \text{in } \Omega_+, \\ u = \phi &\quad \text{on } \partial\Omega_+ \setminus \{x_2 = 0\}, \\ u = 0 &\quad \text{on } \{x_2 = 0\}. \end{aligned}$$

A fairly direct argument shows the following proposition.

PROPOSITION 3.1. *Let u_ϵ and u be as defined above, with u_ϵ extended to all of Ω_+ by setting it to zero on D_ϵ. Then $u_\epsilon \to u$ in $H^1(\Omega_+)$ as $\epsilon \to 0$. Let K denote a compact subset of $\overline{\Omega_+} \setminus \{x_1^2 + (x_2 - 1)^2 = 1\}$. Then $u_\epsilon \to u$ in $C^\infty(K)$ as $\epsilon \to 0$.*

Proof. Let $0 \le \psi \le 1$ be in $C^\infty(\mathbb{R}^2)$ with

$$\psi(y) = 0 \quad \text{for } |y| \le 1 \quad \text{and } \psi(y) = 1 \quad \text{for } |y| \ge 2$$

and define the function v_ϵ by

$$v_\epsilon(x) = u(x)\psi\Big(\frac{x_1}{\epsilon}, \frac{x_2 - \epsilon}{\epsilon}\Big).$$

The function $v_\epsilon \in H^1(\Omega_+)$ satisfies

$$\begin{aligned} v_\epsilon = \phi &\quad \text{on } \partial\Omega_+ \setminus \{x_2 = 0\}, \\ v_\epsilon = 0 &\quad \text{on } \{x_2 = 0\} \quad \text{and on } \partial D_\epsilon. \end{aligned}$$

Thus, due to Dirichlet's principle,

$$\begin{aligned}\left[\int_{\Omega_+\setminus D_\epsilon} a|\nabla u_\epsilon|^2\,dx\right]^{1/2} &\le \left[\int_{\Omega_+\setminus D_\epsilon} a|\nabla v_\epsilon|^2\,dx\right]^{1/2}\\ &\le \left[\int_{\Omega_+\setminus D_\epsilon} a|\nabla u|^2\,dx\right]^{1/2} + \frac{1}{\epsilon}\left[\int_{\Omega_+\setminus D_\epsilon} a|u|^2|\nabla\psi(\frac{x_1}{\epsilon},\frac{x_2-\epsilon}{\epsilon})|^2\,dx\right]^{1/2}\\ &= \left[\int_{\Omega_+\setminus D_\epsilon} a|\nabla u|^2\,dx\right]^{1/2} + o(1).\end{aligned}$$

For the last estimate we have used the fact that $\nabla\psi(y)$ vanishes for $|y|\ge 2$, and the fact that u is continuous at 0 with $u(0)=0$ (due to the result of DeGiorgi and Nash) to conclude that

$$\left[\int_{\Omega_+\setminus D_\epsilon} a|u|^2|\nabla\psi(\frac{x_1}{\epsilon},\frac{x_2-\epsilon}{\epsilon})|^2\,dx\right]^{1/2} \le C\epsilon \max_{\Omega_+\cap\{x_1^2+(x_2-\epsilon)^2\le 4\epsilon^2\}} |u| = o(\epsilon).$$

Since u_ϵ has been extended to be zero on D_ϵ, and since u is in $H^1(\Omega_+)$, it follows that

$$\int_{\Omega_+} a|\nabla u_\epsilon|^2\,dx \le \int_{\Omega_+} a|\nabla u|^2\,dx + o(1). \tag{5}$$

On the other hand, Dirichlet's principle also gives

$$\int_{\Omega_+} a|\nabla u|^2\,dx \le \int_{\Omega_+} a|\nabla u_\epsilon|^2\,dx. \tag{6}$$

A combination of (5) and (6) yields

$$\int_{\Omega_+} a|\nabla u_\epsilon|^2\,dx = \int_{\Omega_+} a|\nabla u|^2\,dx + o(1). \tag{7}$$

It is easy to see that

$$\int_{\Omega_+} a|\nabla(u_\epsilon - u)|^2\,dx = \int_{\Omega_+} a|\nabla u_\epsilon|^2\,dx - \int_{\Omega_+} a|\nabla u|^2\,dx,$$

and therefore by insertion of (7)

$$\int_{\Omega_+} a|\nabla(u_\epsilon - u)|^2\,dx \to 0 \quad \text{as } \epsilon\to 0.$$

Since $u_\epsilon = u = 0$ on $\{x_2=0\}$, it follows immediately that $u_\epsilon \to u$ in $H^1(\Omega_+)$.

The statement about C^∞ convergence follows from elliptic regularity theory and from the fact that ϕ comes from an odd C^∞ function on all of $\partial\Omega$. We have to exclude the curve $\{x_1^2+(x_2-1)^2=1\}$, since the coefficient a is discontinuous across it. □

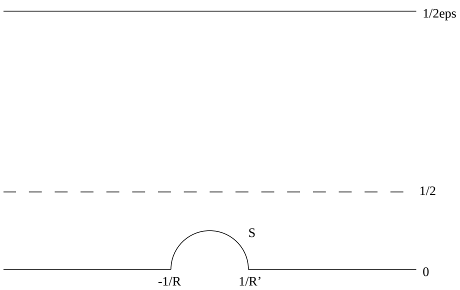


FIG. 7. *The geometric situation in the z-coordinates.*

4. Preliminary estimates. It turns out to be somewhat simpler to work with the variables $(z_1, z_2) = \Phi(x_1, x_2)$ given by the conformal transformation

$$\Phi(x_1, x_2) = \left(\frac{x_1}{x_1^2 + x_2^2}, \frac{x_2}{x_1^2 + x_2^2} \right).$$

The geometric situation is now as illustrated in Figure 7. The "outer" boundary $\partial\Omega \cap \{x_2 > 0\}$ maps to the "inner" boundary S. The circle $\{x_1^2 + (x_2 - 1)^2 = 1\}$ (the fiber) and the circle $\{x_1^2 + (x_2 - \epsilon)^2 = \epsilon^2\}$ (the additional boundary for u_ϵ) map to the horizontal straight lines $z_2 = 1/2$ and $z_2 = 1/2\epsilon$, respectively. The inside of these circles map to the halfplanes $z_2 > 1/2$ and $z_2 > 1/2\epsilon$, respectively. The lower boundary $\{x_2 = 0, -R < x_1 < R'\}$ maps to the straight part of the lower boundary, $\{z_2 = 0, z_1 < -1/R \text{ or } 1/R' < z_1\}$. Since Φ is a conformal mapping, it follows immediately that

$$\int_{\Omega_+} a(x)|\nabla_x v|^2 \, dx = \int_{\Phi(\Omega_+)} A(z)|\nabla_z V|^2 \, dz, \forall v \in H^1(\Omega_+). \tag{8}$$

Here A and V are related to a and v by

$$V(z) = v \circ \Phi^{-1}(z), \quad A(z) = a \circ \Phi^{-1}(z).$$

The transformed solutions $U(z) = u \circ \Phi^{-1}(z)$ and $U_\epsilon(z) = u_\epsilon \circ \Phi^{-1}(z)$ satisfy the differential equations

$$\nabla \cdot A(z)\nabla U = 0 \text{ for } z \in \Phi(\Omega_+)$$

and

$$\nabla \cdot A(z)\nabla U_\epsilon = 0 \text{ for } z \in \Phi(\Omega_+), \quad z_2 < 1/2\epsilon.$$

The transformed conductivity $A(z) = A(z_2)$ has the form

$$A(z) = 1 \text{ for } z_2 < 1/2, \quad A(z) = a_0 \text{ for } z_2 > 1/2.$$

On the "inner" boundary S the functions U and U_ϵ satisfy the boundary conditions

$$U(z) = U_\epsilon(z) = \phi \circ \Phi^{-1}(z).$$

Furthermore, U satisfies

$$U = 0 \text{ on the lower straight boundary } \{z_2 = 0, \quad z_1 < -1/R \text{ or } 1/R' < z_1\},$$

and U_ϵ satisfies

$U_\epsilon = 0$ on the lower straight boundary $\{z_2 = 0, \quad z_1 < -1/R \text{ or } 1/R' < z_1\}$, and on $\{z_2 = 1/2\epsilon\}$.

From the energy identity (8) and the fact that $u, u_\epsilon \in H^1(\Omega_+)$, with $u_\epsilon \to u$, it follows immediately that

(9)
$$\int_{\Phi(\Omega_+)} |\nabla U|^2 \, dz < \infty \quad \text{and} \quad \int_{\Phi(\Omega_+)\cap\{z_2<1/2\epsilon\}} |\nabla U_\epsilon|^2 \, dz < C \text{ independently of } \epsilon.$$

The function $U(z)$ tends to zero as $|z| \to \infty$ (this follows from the fact that u is continuous and has the value 0 at the origin). A simple calculation, using separation of variables, shows that the auxiliary function, U_ϵ, has the expansion

$$U_\epsilon(z) = \sum_{n=1}^{\infty} \beta_{n,\epsilon}\phi_{n,\epsilon}(z_2)e^{-\sqrt{\lambda_{n,\epsilon}}\ z_1} \quad \text{for } z_1 \text{ sufficiently positive,} \tag{10}$$

$$U_\epsilon(z) = \sum_{n=1}^{\infty} \beta'_{n,\epsilon}\phi_{n,\epsilon}(z_2)e^{\sqrt{\lambda_{n,\epsilon}}\ z_1} \quad \text{for } z_1 \text{ sufficiently negative,} \tag{11}$$

where $\lambda_{n,\epsilon} > 0$ and $\phi_{n,\epsilon}(\cdot)$ are the eigenvalues and the eigenvectors of the two point boundary value problem

$$\begin{aligned} -(A(\cdot)\phi')' &= \lambda A(\cdot)\phi \qquad \text{in } (0, 1/2\epsilon), \\ \phi(0) &= \phi(1/2\epsilon) = 0. \end{aligned} \tag{12}$$

Since the coefficient A is bounded from above and is bounded away from zero, it is not difficult to see that

$$d(n\epsilon)^2 \le \lambda_{n,\epsilon} \le D(n\epsilon)^2$$

for some constants $0 < d < D$, independent of n and ϵ. We assume the $\phi_{n,\epsilon}$ are normalized by $\|\phi_{n,\epsilon}\|_{L^2(0,1/2\epsilon)} = 1$. Note that there are no exponentially increasing terms in either of the representations (10) and (11) due to the second inequality from (9). Using the same inequality from (9) (and a standard trace theorem) it follows that there exists $z_1^* > 0$, such that

$$\begin{aligned} \sum_{n=1}^{\infty}(\beta_{n,\epsilon})^2 e^{-2\sqrt{\lambda_{n,\epsilon}}\ z_1^*} &\le C\|U_\epsilon|_{z_1=z_1^*}\|^2_{L^2(0,1/2\epsilon)} \\ &\le C_\epsilon \int_{\Phi(\Omega_+)\cap\{0\le z_1\le z_1^*, 0\le z_2\le 1/2\epsilon\}} |\nabla U_\epsilon|^2 \, dz \\ &< \infty \end{aligned}$$

for any fixed $\epsilon > 0$. Similarly we get that

$$\sum_{n=1}^{\infty}(\beta'_{n,\epsilon})^2 e^{-2\sqrt{\lambda_{n,\epsilon}}\ z_1^*} < \infty$$

for any fixed $\epsilon > 0$. As a consequence of these two bounds and the fact that $d(n\epsilon)^2 \leq \lambda_{n,\epsilon} \leq D(n\epsilon)^2$, it follows immediately that

$$|U_\epsilon(z)| \leq C_\epsilon e^{-c_\epsilon |z_1|}$$

for $|z_1|$ sufficiently large, uniformly in $0 < z_2 < 1/2\epsilon$. That is, for a fixed ϵ, the function U_ϵ converges exponentially to zero as $|z_1| \to \infty$.

Consider the restriction of U_ϵ to the horizontal line $z_2 = 1$ (supposing $\epsilon < 1/2$). By simple integration by parts and Hölder's inequality,

$$\int_{z_2=1} U_\epsilon^2 \, dz_1 \leq C \left(\int_{\Phi(\Omega_+)\cap\{z_2<1\}} \left(\frac{\partial}{\partial z_2} U_\epsilon \right)^2 dz + \int_S U_\epsilon^2 \, ds_z \right).$$

Using the energy bound (9) for U_ϵ and the boundedness of the boundary values of U_ϵ on S, we now obtain

$$\int_{z_2=1} U_\epsilon^2 \, dz_1 \leq C \left(\int_{\Phi(\Omega_+)\cap\{z_2<1\}} |\nabla U_\epsilon|^2 \, dz + \int_S \left(\phi \circ \Phi^{-1}\right)^2 ds_z \right) \;\; \leq \;\; C. \tag{13}$$

The fact that $U_\epsilon(z)$ approaches 0 exponentially fast as $z_1 \to \pm\infty$ along the line $z_2 = 1$ translates into the fact that the function $u_\epsilon(x)$, restricted to the corresponding circle $\{x_1^2 + (x_2 - 1/2)^2 = 1/4\}$, is C^∞, with its value and all the values of its tangential derivatives vanishing at zero. Let v_ϵ denote the solution to

$$\begin{aligned} \triangle v_\epsilon = 0 \quad & \text{in } \{x_1^2 + (x_2 - 1/2)^2 < 1/4\}, \\ v_\epsilon = u_\epsilon \quad & \text{on } \{x_1^2 + (x_2 - 1/2)^2 = 1/4\}. \end{aligned}$$

We introduce one more auxiliary function $V_\epsilon(z)$, the "push-forward" of the function v_ϵ:

$$V_\epsilon(z) = v_\epsilon \circ \Phi^{-1}(z).$$

V_ϵ is defined on the halfplane $\{z_2 > 1\}$. V_ϵ, together with all its derivatives with respect to z_1, converges to zero as $|z| \to \infty$, and it satisfies

$$\triangle V_\epsilon = 0 \quad \text{in } \{z_2 > 1\}, \; V_\epsilon = U_\epsilon \quad \text{on } \{z_2 = 1\}.$$

We have the following representation for V_ϵ:

$$V_\epsilon(z_1, z_2) = \frac{(z_2 - 1)}{\pi} \int_{-\infty}^{\infty} \frac{U_\epsilon(s, 1)}{(z_1 - s)^2 + (z_2 - 1)^2} \, ds,$$

from which it immediately follows that

$$\begin{aligned} |V_\epsilon(z_1, z_2)| &\leq C(z_2 - 1) \|U_\epsilon(\cdot, 1)\|_{L^p(-\infty,\infty)} \| \frac{1}{(z_1 - \cdot)^2 + (z_2 - 1)^2} \|_{L^q(-\infty,\infty)} \\ &\leq C(z_2 - 1)^{-\frac{1}{p}} \|U_\epsilon(\cdot, 1)\|_{L^p(-\infty,\infty)}. \end{aligned} \tag{14}$$

Here $1 \leq p \leq \infty$, and $\frac{1}{p} + \frac{1}{q} = 1$. Note that, since it is continuous as a function of z_1 and since it decreases exponentially to zero as $z_1 \to \infty$, $U_\epsilon(z_1, 1)$ belongs to $L^p(-\infty, \infty)$ for all $1 \leq p \leq \infty$. For any integer $k \geq 1$, we similarly get

$$\left| \left(\frac{\partial}{\partial z_1} \right)^k V_\epsilon(z_1, z_2) \right| \leq C(z_2 - 1)^{-k-\frac{1}{p}} \|U_\epsilon(\cdot, 1)\|_{L^p(-\infty,\infty)}. \tag{15}$$

The estimates (13), (14), and (15) (the latter two for $p = 2$) lead to the following lemma.

LEMMA 4.1. *Let $0 < \alpha$. Then there exists a constant C, independent of ϵ, such that*

$$|U_\epsilon(z)| \le C|z|^{-1/2}, \quad \left|\frac{\partial}{\partial z_1}U_\epsilon(z)\right| \le C|z|^{-3/2}, \quad \textit{and} \ \left|\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon(z)\right| \le C|z|^{-5/2}$$

for $z \in \{\max(2, \alpha|z_1|) \le z_2 \le 1/2\epsilon\}$.

Proof. Let V_ϵ be as above. The function $U_\epsilon - V_\epsilon$ satisfies

$$\begin{aligned} \triangle(U_\epsilon - V_\epsilon) = 0 \quad &\text{in } \{1 < z_2 < 1/2\epsilon\}, \\ U_\epsilon - V_\epsilon = 0 \quad &\text{on } \{z_2 = 1\}, \\ |U_\epsilon - V_\epsilon| \le C\sqrt{\epsilon} \quad &\text{on } \{z_2 = 1/2\epsilon\}. \end{aligned} \tag{16}$$

The estimate of $U_\epsilon - V_\epsilon$ on $\{z_2 = 1/2\epsilon\}$ follows from (13) and (14) with $p = 2$. For any fixed ϵ we also have that $(U_\epsilon - V_\epsilon)(z) \to 0$ as $|z_1| \to \infty$ (uniformly on $1 \le z_2 \le 1/2\epsilon$). An application of the maximum principle to (16) now yields

$$|(U_\epsilon - V_\epsilon)(z)| \le C\sqrt{\epsilon} \quad \text{on } \{1 \le z_2 \le 1/2\epsilon\},$$

and therefore, based on (13) and (14) with $p = 2$

$$|U_\epsilon(z)| \le C(\sqrt{\epsilon} + z_2^{-1/2}) \quad \text{on } \{2 \le z_2 \le 1/2\epsilon\}.$$

It follows from this that

$$|U_\epsilon(z)| \le C|z|^{-1/2} \quad \text{on } \{\max(2, \alpha|z_1|) \le z_2 \le 1/2\epsilon\},$$

as desired. A similar argument may be used to derive the desired estimates for $\frac{\partial}{\partial z_1}U_\epsilon$ and $\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon$, the only difference being that one applies (15) with $k = 1$ and 2 in place of (14). ☐

The following result will prove convenient both here and later.

LEMMA 4.2. *Let $0 < M_0$, $0 < \gamma$ and $0 < \alpha$, with $\arctan\alpha < \frac{\pi}{2(\gamma+1)}$. Suppose W_ϵ is continuous on $\{M_0 \le z_1, \ \ 0 \le z_2 \le \frac{1}{2\epsilon}\}$, continuously differentiable on each of the sets $\{M_0 \le z_1, \frac{1}{2} \le z_2 \le \frac{1}{2\epsilon}\}$ and $\{M_0 \le z_1, 0 \le z_2 \le \frac{1}{2}\}$, and satisfies, in a weak sense,*

$$\begin{aligned} \nabla \cdot A(z)\nabla W_\epsilon = 0 \quad &\textit{in } \{M_0 < z_1, \ \ 0 < z_2 < 1/2\epsilon\} \\ W_\epsilon = 0 \quad &\textit{on } \{M_0 \le z_1, \ \ z_2 = 1/2\epsilon\} \textit{ and } \{M_0 \le z_1, z_2 = 0\}. \end{aligned}$$

Suppose furthermore that $|W_\epsilon(M_0, z_2)| \le K_0$ for $0 \le z_2 \le \alpha M_0$, and that, for fixed ϵ, $W_\epsilon(z)z_1^\gamma \to 0$ as $z_1 \to \infty$, uniformly on $0 \le z_2 \le 1/2\epsilon$. Then

$$|W_\epsilon(z)| \le C_0|z|^{-\gamma} \textit{ on } \{z_2 = \alpha z_1, \ \ \alpha M_0 \le z_2 \le 1/2\epsilon\}$$

implies that

$$|W_\epsilon(z)| \le C_0'|z|^{-\gamma} \textit{ on } \{M_0 \le z_1, \ \ 0 \le z_2 \le \min(\alpha z_1, \ 1/2\epsilon)\},$$

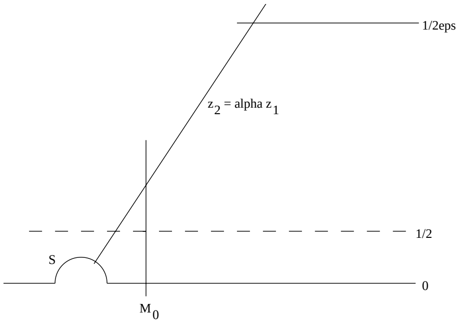


FIG. 8. *The semi-infinite strip. The dashed line represents the conductivity discontinuity at* $z_2 = 1/2$. *The "inner" boundary* S *is illustrated as a semicircle.*

with C_0' *depending on* C_0, M_0, K_0, α, *and* γ *but otherwise independent of* W_ϵ *(and* ϵ*).*

Proof. The proof differs slightly depending on whether $a_0 < 1$ or $a_0 > 1$. We start by considering the case $a_0 < 1$. Let $\mu(z)$ denote the function

$$\mu(z) = r^{-\gamma} \sin \gamma(\theta + \delta),$$

where (r, θ) denotes polar coordinates centered at the origin. Let P_ϵ denote the semi-infinite strip bounded by the four lines $\{z_2 = 0\}$, $\{z_1 = M_0\}$, $\{z_2 = \alpha z_1\}$, and $\{z_2 = 1/2\epsilon\}$ (see Figure 8).

The function μ is clearly harmonic in P_ϵ. We also calculate

$$\begin{aligned}\frac{\partial \mu}{\partial z_2} = \sin\theta \frac{\partial \mu}{\partial r} + \frac{\cos\theta}{r}\frac{\partial \mu}{\partial \theta} &= -\gamma r^{-\gamma-1}(\sin\theta \sin\gamma(\theta+\delta) - \cos\theta\cos\gamma(\theta+\delta)) \\ &= \gamma r^{-\gamma-1}\cos((\gamma+1)\theta + \gamma\delta).\end{aligned}$$

Due to the condition $\arctan \alpha < \frac{\pi}{2(\gamma+1)}$, it follows that $0 \le \theta \le \arctan\alpha < \frac{\pi}{2(\gamma+1)}$ on $\overline{P_\epsilon}$. By selecting $0 < \delta$ sufficiently small we may thus obtain

$$0 < (\gamma+1)\theta + \gamma\delta < \pi/2 \quad \text{on } \overline{P_\epsilon}.$$

It follows immediately that

$$\mu(z) > 0 \quad \text{for } z \in \overline{P_\epsilon} \quad \text{ and } \quad \frac{\partial \mu}{\partial z_2}(z) > 0 \quad \text{for } z \in P_\epsilon. \tag{17}$$

Now consider the function

$$w_\epsilon(z) = \frac{W_\epsilon(z)}{\mu(z)}.$$

A simple calculation gives that

$$\triangle w_\epsilon + 2\frac{1}{\mu}\nabla\mu \cdot \nabla w_\epsilon = 0 \quad \text{in } P_\epsilon \setminus \{z_2 = 1/2\} \tag{18}$$

and

$$a_0 \frac{\partial w_\epsilon}{\partial z_2}^+ - \frac{\partial w_\epsilon}{\partial z_2}^- = (1-a_0)\frac{1}{\mu}\frac{\partial \mu}{\partial z_2} w_\epsilon \quad \text{on the half-line } \{z_2 = 1/2\} \cap P_\epsilon\ . \tag{19}$$

Equation (18), the fact that w_ϵ attains the value 0 on ∂P_ϵ, and the fact that $w_\epsilon(z) \to 0$ as $z_1 \to \infty$ imply that w_ϵ attains it extremal values (min and max) on ∂P_ϵ or on the half-line $\{z_2 = 1/2\} \cap P_\epsilon$. The condition (19) rules out the possibility that an extremal value can be attained on $\{z_2 = 1/2\} \cap P_\epsilon$ unless w_ϵ is constant (= 0): if a maximum was attained at $z_0 \in \{z_2 = 1/2\} \cap P_\epsilon$ then $w_\epsilon(z_0) \geq 0$ and thus, according to (17) and (19),

$$a_0 \frac{\partial w_\epsilon}{\partial z_2}^+ (z_0) - \frac{\partial w_\epsilon}{\partial z_2}^- (z_0) \geq 0.$$

(Remember we are in the case $a_0 < 1$.) However, Hopf's version of the maximum principle asserts that if w_ϵ is not constant (= 0) then $\frac{\partial w_\epsilon}{\partial z_2}^+(z_0) \leq 0$ and $\frac{\partial w_\epsilon}{\partial z_2}^-(z_0) \geq 0$, and at least one of these values is nonzero. Consequently

$$a_0 \frac{\partial w_\epsilon}{\partial z_2}^+ (z_0) - \frac{\partial w_\epsilon}{\partial z_2}^- (z_0) < 0,$$

and this represents a contradiction. Corresponding to a minimum we would have $w_\epsilon(z_0) \leq 0$, and the same argument as above would lead to a contradiction (unless $w_\epsilon = 0$).

We may therefore conclude that the extremal values of w_ϵ are always attained on ∂P_ϵ. Let d_0 denote the constant

$$d_0 = \sin \gamma\delta.$$

It follows that $0 < d_0$ and that d_0 only depends on γ and δ. It is quite easy to see that

$$|w_\epsilon(z)| \leq K_0 d_0^{-1} M_0^\gamma (1+\alpha^2)^{\gamma/2}$$

on $\{z_1 = M_0, \ \ 0 \leq z_2 \leq \alpha M_0\}$ and that

$$|w_\epsilon(z)| \leq C_0 d_0^{-1}$$

on $\{z_2 = \alpha z_1,\ \alpha M_0 \leq z_2 \leq 1/2\epsilon\}$. Together these two estimates show that

$$|w_\epsilon(z)| \leq C_0' \quad \text{on } \partial P_\epsilon,$$

with $C_0' = d_0^{-1} \max(C_0, K_0 M_0^\gamma (1+\alpha^2)^{\gamma/2})$. Since the extremal values of w_ϵ are attained on the boundary of P_ϵ, we immediately conclude that $|w_\epsilon(z)| \leq C_0'$ in P_ϵ, and thus

$$|W_\epsilon(z)| \leq C_0' |z|^{-\gamma} \quad \text{in } P_\epsilon,$$

exactly as desired (for $a_0 < 1$).

In the case $a_0 > 1$ we introduce the function $\mu = r^{-\gamma} \cos\gamma\theta$. We now calculate

$$\begin{aligned}\frac{\partial \mu}{\partial z_2} = \sin\theta \frac{\partial \mu}{\partial r} + \frac{\cos\theta}{r}\frac{\partial \mu}{\partial \theta} &= -\gamma r^{-\gamma-1}(\sin\theta\cos\gamma\theta + \cos\theta\sin\gamma\theta)\\ &= -\gamma r^{-\gamma-1}\sin(\gamma+1)\theta.\end{aligned}$$

Therefore

$$\mu(z) > 0 \quad \text{for } z \in \overline{P_\epsilon} \quad \text{and} \quad \frac{\partial \mu}{\partial z_2} < 0 \quad \text{for } z \in P_\epsilon.$$

The argument from before works in an identical fashion with this function, since the signs of $1 - a_0$ and $\frac{\partial}{\partial z_2}\mu$ have both changed, so that $\frac{1}{\mu}(1 - a_0)\frac{\partial}{\partial z_2}\mu$ stays positive. The constant d_0 gets replaced by

$$d_0' = \cos \gamma\theta_\alpha > 0,$$

with $\theta_\alpha = \arctan \alpha$. This shows that

$$|W_\epsilon(z)| \leq C_0'|z|^{-\gamma} \quad \text{in } P_\epsilon,$$

exactly as desired (for $a_0 > 1$ as well). □

Based on the two previous lemmas it is now fairly simple to prove the following proposition.

Proposition 4.3. *There exists a constant C, independent of ϵ, such that the functions U_ϵ satisfy*

$$|U_\epsilon(z)| \leq C|z|^{-1/2}, \quad \left|\frac{\partial}{\partial z_1}U_\epsilon(z)\right| \leq C|z|^{-3/2}, \text{ and } \left|\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon(z)\right| \leq C|z|^{-5/2}$$

for $z \in \Phi(\Omega_+) \cap \{z_2 < 1/2\epsilon\}$.

Proof. Select $0 < \alpha$ so that $\arctan \alpha < \pi/3$, and select $0 < M$ so that $2 < \alpha M$, and so that the line $\{z_1 = M\}$ does not intersect the "inner" boundary S. Elliptic regularity results (in combination with the uniform energy bound for U_ϵ) easily give

$$|U_\epsilon(M, z_2)| \leq K \quad \text{for } 0 \leq z_2 \leq \alpha M,$$

with K independent of ϵ. We also have that $U_\epsilon(z)z_1^{1/2} \to 0$ as $z_1 \to \infty$ (uniformly on $0 \leq z_2 \leq 1/2\epsilon$). Indeed, U_ϵ decreases exponentially as $z_1 \to \infty$. From Lemma 4.1 we know that

$$|U_\epsilon(z)| \leq C|z|^{-1/2} \quad \text{on } \{z_2 = \alpha z_1,\ \alpha M \leq z_2 \leq 1/2\epsilon\}.$$

Application of Lemma 4.2 with $\gamma = 1/2$ now gives

$$|U_\epsilon(z)| \leq C'|z|^{-1/2} \quad \text{on } \{M \leq z_1,\ 0 \leq z_2 \leq \min(\alpha z_1,\ 1/2\epsilon)\},$$

with C' independent of ϵ. For $0 \leq z_1$, $\max(\alpha M, \alpha z_1) \leq z_2 \leq 1/2\epsilon$ it follows immediately from Lemma 4.1 that $|U_\epsilon(z)| \leq C|z|^{-1/2}$. For z outside S, $0 \leq z_1 \leq M$ and $0 \leq z_2 \leq \alpha M$ (the remainder of $\Phi(\Omega_+) \cap \{0 \leq z_1,\ z_2 \leq 1/2\epsilon\}$) elliptic regularity results yield that $|U_\epsilon(z)| \leq C \leq C|z|^{-1/2}$. In summary, we have thus verified that

$$|U_\epsilon(z)| \leq C|z|^{-1/2} \quad \text{in } \Phi(\Omega_+) \cap \{0 \leq z_1,\ z_2 < 1/2\epsilon\}.$$

A similar argument (e.g., using $U_\epsilon(-z_1, z_2)$ in place of $U(z_1, z_2)$) proves the same estimate in $\Phi(\Omega_+) \cap \{z_1 \leq 0,\ z_2 < 1/2\epsilon\}$, thus completing the proof of the first assertion of this proposition. Almost identical arguments with $\gamma = 3/2$ and $\gamma = 5/2$ lead to the desired estimates for $\frac{\partial}{\partial z_1}U_\epsilon$ and $\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon$, respectively. Note that these functions also solve the type of boundary value problem required in Lemma 4.2. □

Since $U_\epsilon \to U$, $\frac{\partial}{\partial z_1}U_\epsilon \to \frac{\partial}{\partial z_1}U$ and $\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon \to \left(\frac{\partial}{\partial z_1}\right)^2 U$ pointwise inside $\Phi(\Omega_+) \setminus \{z_2 = 1/2\}$ (Proposition 3.1) and since U, $\frac{\partial}{\partial z_1}U$, and $\left(\frac{\partial}{\partial z_1}\right)^2 U$ are all continuous in $\Phi(\Omega_+)$, we derive from the above proposition the following corollary.

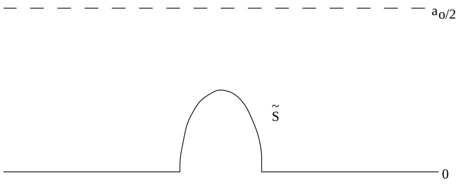


Fig. 9. *The stretched geometry. The dashed line represents the location of the discontinuity of the coefficient $\tilde{A}$.*

Corollary 4.4. *There exists a constant C such that*

$$|U(z)| \leq C|z|^{-1/2}, \quad \left|\frac{\partial}{\partial z_1}U(z)\right| \leq C|z|^{-3/2}, \text{ and } \left|\left(\frac{\partial}{\partial z_1}\right)^2 U(z)\right| \leq C|z|^{-5/2}$$

for $z \in \Phi(\Omega_+)$.

Based on the use of Proposition 4.3 we are able to establish improved estimates for U_ϵ and U that immediately lead to a proof of our main theorem.

5. The improved estimates. We extend the function A to all of the halfplane $0 < z_2$ by setting it to 1 in the domain bounded by $S = \Phi(\partial\Omega \cap \{x_2 > 0\})$ and $\{z_2 = 0\}$, and we introduce the new variables $\tilde{z} = (\tilde{z}_1, \tilde{z}_2)$ as follows:

$$\tilde{z}_1 = z_1, \quad \tilde{z}_2 = a_0 \int_0^{z_2} \frac{1}{A(s)}\, ds = \begin{cases} z_2 + \frac{a_0 - 1}{2}, & z_2 > 1/2, \\ a_0 z_2, & z_2 < 1/2. \end{cases} \tag{20}$$

The transformed function $\tilde{U}(\tilde{z}) = U(z)$ solves

$$\left[\left(\frac{\tilde{A}(\tilde{z}_2)}{a_0}\right)^2 \left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 + \left(\frac{\partial}{\partial \tilde{z}_2}\right)^2\right] \tilde{U} = 0 \quad \text{for } \tilde{z} \in \tilde{\Phi}(\Omega_+),$$

with the domain $\tilde{\Phi}(\Omega_+)$ defined by

$$\tilde{z} \in \tilde{\Phi}(\Omega_+) \quad \text{if and only if } z \in \Phi(\Omega_+).$$

See Figure 9. We have illustrated the stretched "inner" boundary $\tilde{S}$ as the upper half of an ellipse.

The function $\tilde{A}$ is defined by $\tilde{A}(\tilde{z}_2) = A(z_2)$, where $\tilde{z}_2$ and z_2 are related by the second formula in (20). From the above definitions it follows immediately that

$$\triangle_{\tilde{z}} \tilde{U} = 0 \quad \text{for } a_0/2 < \tilde{z}_2$$

and

$$\triangle_{\tilde{z}} \tilde{U} = \left(1 - \frac{1}{a_0^2}\right) \left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{U} \quad \text{for } \tilde{z} \in \tilde{\Phi}(\Omega_+) \cap \{\tilde{z}_2 < a_0/2\}.$$

Similarly,

$$\triangle_{\tilde{z}} \tilde{U}_\epsilon = 0 \quad \text{for } a_0/2 < \tilde{z}_2 < \frac{1 + \epsilon(a_0 - 1)}{2\epsilon}$$

and

$$\triangle_{\tilde{z}}\tilde{U}_\epsilon = (1-\frac{1}{a_0^2})\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{U}_\epsilon \quad \text{for } \tilde{z} \in \tilde{\Phi}(\Omega_+) \cap \{\tilde{z}_2 < a_0/2\}.$$

Let $\tilde{E}$ denote the domain bounded by the inner boundary $\tilde{S}$ and $\{\tilde{z}_2 = 0\}$. In Figure 9 it is represented by the inside of the half-ellipse. We now extend both $\tilde{U}$ (and $\tilde{U}_\epsilon$) to all of the halfplane $0 < \tilde{z}_2$ (the strip $0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}$) in such a way that the extensions are zero on $\tilde{z}_2 = 0$ and are $C^{2,\beta}$ bounded in a neighborhood of $\tilde{E}$ (independently of ϵ). This may be done since, near $\tilde{S}$, $\tilde{U}$ and $\tilde{U}_\epsilon$ are C^∞ and uniformly bounded in all C^k norms (due to elliptic regularity results and the uniform energy estimate). As a consequence we get that

$$\begin{aligned}
&\triangle_{\tilde{z}}\tilde{U} = 0 \quad \text{for } a_0/2 < \tilde{z}_2,\\
&\triangle_{\tilde{z}}\tilde{U} = \left(1-\frac{1}{a_0^2}\right)\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{U} + f(\tilde{z}) \quad \text{for } 0 < \tilde{z}_2 < a_0/2,
\end{aligned}$$

where f is uniformly bounded and supported in the closure of $\tilde{E}$. We also get that

$$\begin{aligned}
&\triangle_{\tilde{z}}\tilde{U}_\epsilon = 0 \quad \text{for } a_0/2 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon},\\
&\triangle_{\tilde{z}}\tilde{U}_\epsilon = \left(1-\frac{1}{a_0^2}\right)\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{U}_\epsilon + f_\epsilon(\tilde{z}) \quad \text{for } 0 < \tilde{z}_2 < a_0/2,
\end{aligned}$$

where the f_ϵ are uniformly bounded (independently of ϵ) and supported in the closure of $\tilde{E}$. The functions $\tilde{U}$, $\tilde{U}_\epsilon$ and the gradients $\nabla_{\tilde{z}}\tilde{U}$, $\nabla_{\tilde{z}}\tilde{U}_\epsilon$ are continuous across the line $\tilde{z}_2 = a_0/2$, so the above piecewise formulas entirely describe (the distributions) $\triangle_{\tilde{z}}\tilde{U}$ and $\triangle_{\tilde{z}}\tilde{U}_\epsilon$. Let g_ϵ denote the function $g_\epsilon = (1-\frac{1}{a_0^2})\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2\tilde{U}_\epsilon + f_\epsilon$ (in $\tilde{E}$) and define

$$\begin{aligned}
\tilde{V}_\epsilon^*(\tilde{z}) = \frac{a_0^2-1}{4\pi a_0^2}\int_{\{(s,t)\,:\,0<t<a_0/2\}\setminus\tilde{E}} \Bigg[&\log(|\tilde{z}_1-s|^2+|\tilde{z}_2-t|^2)\\
&-\log(|\tilde{z}_1-s|^2+|\tilde{z}_2+t|^2)\Bigg]\left(\frac{\partial}{\partial s}\right)^2\tilde{U}_\epsilon(s,t)\,dsdt\\
+\frac{1}{4\pi}\int_{\tilde{E}}\Bigg[&\log(|\tilde{z}_1-s|^2+|\tilde{z}_2-t|^2)\\
&-\log(|\tilde{z}_1-s|^2+|\tilde{z}_2+t|^2)\Bigg]g_\epsilon(s,t)\,dsdt.
\end{aligned}$$

It is not difficult to see that $\tilde{V}_\epsilon^*$ satisfies

$$\begin{aligned}
&\triangle_{\tilde{z}}\tilde{V}_\epsilon^* = 0 \quad \text{for } a_0/2 < \tilde{z}_2,\\
&\triangle_{\tilde{z}}\tilde{V}_\epsilon^* = \left(1-\frac{1}{a_0^2}\right)\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{U}_\epsilon + f_\epsilon(\tilde{z}) \quad \text{for } 0 < \tilde{z}_2 < a_0/2, \quad \text{and}\\
&\tilde{V}_\epsilon^* = 0 \quad \text{at } \tilde{z}_2 = 0.
\end{aligned}$$

$\tilde{V}_\epsilon^*$ and $\nabla_{\tilde{z}}\tilde{V}_\epsilon^*$ are continuous across the line $\tilde{z}_2 = a_0/2$, so the above piecewise formula entirely describes (the distribution) $\triangle_{\tilde{z}}V_\epsilon^*$. The second order derivative $\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2\tilde{V}_\epsilon^*$ is also continuous across the line $\tilde{z}_2 = a_0/2$, but the second order derivative $\left(\frac{\partial}{\partial \tilde{z}_2}\right)^2 \tilde{V}_\epsilon^*$ is in general discontinuous.

Lemma 5.1. *For any fixed ϵ we have that*

$$|\tilde{V}_\epsilon^*(\tilde{z})| \to 0, \qquad \left|\frac{\partial}{\partial \tilde{z}_1}\tilde{V}_\epsilon^*(\tilde{z})\right| \to 0, \quad \textit{and} \ \left|\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{V}_\epsilon^*(\tilde{z})\right| \to 0$$

as $|\tilde{z}_1| \to \infty$, uniformly for $\tilde{z} \in \{0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}\}$.

Proof. Since the function g_ϵ is uniformly bounded independently of ϵ, and since

$$\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 - t|^2) - \log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + t|^2) = -2t\frac{2(\tilde{z}_2 + \theta)}{|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + \theta|^2},$$

where $-a_0/2 < \theta < a_0/2$ (when $0 < t < a_0/2$), it follows immediately that the last integral in the definition of $\tilde{V}_\epsilon^*$,

$$\frac{1}{4\pi}\int_{\tilde{E}} \left[\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 - t|^2) - \log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + t|^2)\right] g_\epsilon(s,t)\, dsdt,$$

converges to 0 as $|\tilde{z}_1| \to \infty$ (uniformly for $0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}$). Let $\tilde{E}^c$ denote the set

$$\tilde{E}^c = \{(s,t) \ : \ 0 < t < a_0/2\} \setminus \tilde{E}.$$

The first integral in the definition of $\tilde{V}_\epsilon^*$ may be bounded by

$$\begin{aligned} C\int_{\tilde{E}^c \cap \{|\tilde{z}_1 - s| < |\tilde{z}_1|/2\}} \Big[\big|\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 - t|^2)\big| & \\ + \big|\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + t|^2)\big| \Big] & \left|\left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)\right| dsdt \\ +C\int_{\tilde{E}^c \cap \{|\tilde{z}_1 - s| \geq |\tilde{z}_1|/2\}} t\frac{|\tilde{z}_2 + \theta|}{|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + \theta|^2} & |\left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)|\, dsdt. \end{aligned} \tag{21}$$

As $|\tilde{z}_1| \to \infty$ (for $0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}$) the second integral in (21) is bounded by

$$C_\epsilon \frac{1}{|\tilde{z}_1|^2}\int_{\tilde{E}^c} |\left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)|\, dsdt,$$

which clearly approaches 0 (the integral in this estimate is uniformly bounded in ϵ according to Proposition 4.3). For $0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}$, $0 < t < a_0/2$ we have

$$|\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 - t|^2)| \leq |\log(|\tilde{z}_1 - s|^2)| + K_\epsilon$$

and

$$|\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + t|^2)| \leq |\log(|\tilde{z}_1 - s|^2)| + K_\epsilon,$$

so that the first integral in (21) becomes bounded by

$$\begin{aligned} C_\epsilon \int_{\tilde{E}^c \cap \{|\tilde{z}_1 - s| < |\tilde{z}_1|/2\}} (|\log(|\tilde{z}_1 - s|^2)| + 1)\left|\left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)\right| dsdt & \\ \leq C_\epsilon e^{-c_\epsilon|\tilde{z}_1|}\int_{\tilde{E}^c \cap \{|\tilde{z}_1 - s| < |\tilde{z}_1|/2\}} (|\log(|\tilde{z}_1 - s|^2)| + 1)\, dsdt & \\ \leq C_\epsilon e^{-c_\epsilon|\tilde{z}_1|}|\tilde{z}_1|\,|\log(|\tilde{z}_1|^2)| \quad \text{as } |\tilde{z}_1| \to \infty. & \end{aligned}$$

Here we used that $|\tilde{z}_1 - s| < |\tilde{z}_1|/2 \Rightarrow |s| > |\tilde{z}_1|/2$, and that $\left| \left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)\right|$ decreases exponentially in $|s|$ (uniformly in t, for fixed ϵ). This proves that the first integral in the definition of $\tilde{V}_\epsilon^*$ converges to 0 as $|\tilde{z}_1| \to \infty$ (uniformly for $0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}$) and it thus verifies the asymptotic statement concerning $\tilde{V}_\epsilon^*$. For the first and second order derivatives of $\tilde{V}_\epsilon^*$ with respect to $\tilde{z}_1$ we write (for $|\tilde{z}_1|$ large)

$$\begin{aligned}\left(\frac{\partial}{\partial \tilde{z}_1}\right)^k \tilde{V}_\epsilon^*(\tilde{z}) = & \frac{a_0^2-1}{4\pi a_0^2} \int_{\{(s,t)\ :\ 0<t<a_0/2\}\setminus \tilde{E}} \Bigg[\log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 - t|^2) \\ & - \log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + t|^2)\Bigg] \left(\frac{\partial}{\partial s}\right)^{2+k} \tilde{U}_\epsilon(s,t)\, dsdt \\ & + \frac{1}{4\pi} \int_{\tilde{E}} \Bigg[\left(\frac{\partial}{\partial \tilde{z}_1}\right)^k \log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 - t|^2) \\ & - \left(\frac{\partial}{\partial \tilde{z}_1}\right)^k \log(|\tilde{z}_1 - s|^2 + |\tilde{z}_2 + t|^2)\Bigg] g_\epsilon(s,t)\, dsdt\end{aligned}$$

and apply an argument very similar to that above. □

Let $H(x,y)$ denote the function

$$H(x,y) = \frac{2(y^2 - x^2)}{(x^2+y^2)^2}.$$

For $a_0/2 < \tilde{z}_2$ a fairly straightforward computation gives

$$\begin{aligned}\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{V}_\epsilon^*(\tilde{z}_1, \tilde{z}_2) = \frac{a_0^2-1}{4\pi a_0^2} \int_{\{(s,t)\ :\ 0<t<a_0/2\}\setminus \tilde{E}} & \big[H(\tilde{z}_1 - s, \tilde{z}_2 - t) \\ & - H(\tilde{z}_1 - s, \tilde{z}_2 + t)\big] \left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)\, dsdt \\ + \frac{1}{4\pi} \int_{\tilde{E}} \left[H(\tilde{z}_1 - s, \tilde{z}_2 - t) - H(\tilde{z}_1 - s, \tilde{z}_2 + t)\right] & g_\epsilon(s,t)\, dsdt.\end{aligned}$$

By use of Taylor's formula we now get

$$\begin{aligned}& \left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{V}_\epsilon^*(\tilde{z}_1, \tilde{z}_2) \\ & \quad = -\frac{a_0^2-1}{4\pi a_0^2} \int_{\{(s,t)\ :\ 0<t<a_0/2\}\setminus \tilde{E}} 2t\ H_y(\tilde{z}_1 - s, \tilde{z}_2 + \theta) \left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)\, dsdt \\ & \qquad - \frac{1}{4\pi} \int_{\tilde{E}} 2t\ H_y(\tilde{z}_1 - s, \tilde{z}_2 + \theta) g_\epsilon(s,t)\, dsdt, \end{aligned} \tag{22}$$

where θ lies between $-a_0/2$ and $a_0/2$ (and depends on $\tilde{z}_1 - s$, $\tilde{z}_2$ and t). The derivative H_y is given by

$$|H_y(x,y)| = \left| \frac{y(12x^2 - 4y^2)}{(x^2+y^2)^3} \right| \le C y^{-3}, \quad 1 < y. \tag{23}$$

From Proposition 4.3 we know that

$$\left|\left(\frac{\partial}{\partial s}\right)^2 \tilde{U}_\epsilon(s,t)\right| \le C(s^2+t^2)^{-5/4}$$

for $(s,t) \in \tilde{E}^c$. We also know that g_ϵ is uniformly bounded on $\tilde{E}$, independently of ϵ. Combining these two facts with (22) and (23) we now conclude that

$$\left|\left(\frac{\partial}{\partial \tilde{z}_1}\right)^2 \tilde{V}^*_\epsilon(\tilde{z}_1,\tilde{z}_2)\right| \le C\tilde{z}_2^{-3}, \quad 1+a_0/2 < \tilde{z}_2, \tag{24}$$

with C independent of ϵ. By an entirely similar argument (taking just one derivative, or no derivative at all) we obtain the estimates

$$\left|\frac{\partial}{\partial \tilde{z}_1}\tilde{V}^*_\epsilon(\tilde{z}_1,\tilde{z}_2)\right| \le C\tilde{z}_2^{-2} \quad \text{and} \quad \left|\tilde{V}^*_\epsilon(\tilde{z}_1,\tilde{z}_2)\right| \le C\tilde{z}_2^{-1} \quad \text{for } 1+a_0/2 < \tilde{z}_2. \tag{25}$$

The estimates (24) and (25) are stronger than the corresponding estimates (14) and (15) (with $p=2$) by a factor of $\tilde{z}_2^{-1/2}$. Not surprisingly, these estimates for $\tilde{V}^*_\epsilon$ lead to improvements of the results of Lemma 4.1 by a factor of $|z|^{-1/2}$.

Lemma 5.2. *Let $0<\alpha$. Then there exists C, independent of ϵ, such that*

$$|U_\epsilon(z)| \le C|z|^{-1}, \quad \left|\frac{\partial}{\partial z_1}U_\epsilon(z)\right| \le C|z|^{-2}, \quad \textit{and} \quad \left|\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon(z)\right| \le C|z|^{-3}$$

for $z \in \{\max(2,\alpha|z_1|) \le z_2 \le 1/2\epsilon\}$.

Proof. The function $\tilde{U}_\epsilon - \tilde{V}^*_\epsilon$ satisfies

$$\begin{aligned} \triangle(\tilde{U}_\epsilon - \tilde{V}^*_\epsilon) &= 0 \quad \text{for } 0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}, \\ \tilde{U}_\epsilon - \tilde{V}^*_\epsilon &= 0 \quad \text{at } \tilde{z}_2 = 0, \\ \tilde{U}_\epsilon - \tilde{V}^*_\epsilon &= O(\epsilon) \quad \text{at } \tilde{z}_2 = \frac{1+\epsilon(a_0-1)}{2\epsilon}, \end{aligned}$$

and for any fixed ϵ, $\tilde{U}_\epsilon - \tilde{V}^*_\epsilon \to 0$ as $|\tilde{z}_1| \to \infty$ (uniformly for $0 < \tilde{z}_2 < \frac{1+\epsilon(a_0-1)}{2\epsilon}$). The desired estimate for U_ϵ follows by an application of the maximum principle just as in the proof of Lemma 4.1 (possibly with a smaller coefficient α') and then a return to the z coordinates. The estimates for $\frac{\partial}{\partial z_1}U_\epsilon$ and $\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon$ follow in a completely similar manner. □

Based on Lemma 4.2 and Lemma 5.2 it is now possible to establish the following proposition.

Proposition 5.3. *There exists a constant C, independent of ϵ, such that the functions U_ϵ satisfy*

$$|U_\epsilon(z)| \le C|z|^{-1}, \quad \left|\frac{\partial}{\partial z_1}U_\epsilon(z)\right| \le C|z|^{-2}, \; \textit{and} \; \left|\left(\frac{\partial}{\partial z_1}\right)^2 U_\epsilon(z)\right| \le C|z|^{-3}$$

for $z \in \Phi(\Omega_+) \cap \{z_2 < 1/2\epsilon\}$.

Proof. The proof is exactly the same as that of Proposition 4.3. □

Taking the limit $\epsilon \to 0$, we conclude just as before.

Corollary 5.4. *There exists a constant C, such that*

$$|U(z)| \le C|z|^{-1}, \quad \left|\frac{\partial}{\partial z_1}U(z)\right| \le C|z|^{-2}, \ \textit{and} \ \left|\left(\frac{\partial}{\partial z_1}\right)^2 U(z)\right| \le C|z|^{-3}$$

for $z \in \Phi(\Omega_+)$.

6. The proof of the main theorem. Going back to the function $\tilde{U}$, one can check that it has the representation formula

$$\begin{aligned}\tilde{U}(\tilde{z}) = \frac{a_0^2-1}{4\pi a_0^2}\int_{\{(s,t)\ :\ 0<t<a_0/2\}\setminus\tilde{E}} \Bigg[\log(|\tilde{z}_1-s|^2+|\tilde{z}_2-t|^2) \\ -\log(|\tilde{z}_1-s|^2+|\tilde{z}_2+t|^2)\Bigg]\left(\frac{\partial}{\partial s}\right)^2\tilde{U}(s,t)\,dsdt \\ +\frac{1}{4\pi}\int_{\tilde{E}}\Bigg[\log(|\tilde{z}_1-s|^2+|\tilde{z}_2-t|^2) \\ -\log(|\tilde{z}_1-s|^2+|\tilde{z}_2+t|^2)\Bigg]g(s,t)\,dsdt,\end{aligned} \tag{26}$$

where g denotes the function

$$g = \left(1-\frac{1}{a_0^2}\right)\left(\frac{\partial}{\partial\tilde{z}_1}\right)^2\tilde{U}+f$$

in $\tilde{E}$. To see this, it suffices to notice that the right hand side of (26) satisfies the same equation and the same boundary condition as $\tilde{U}$ and to verify that it also converges to zero as $|\tilde{z}|$ tends to infinity. The fact that the two integrals in (26) converge to zero for $0 < \tilde{z}_2 < K$ as $|\tilde{z}_1| \to \infty$ follows from an argument very similar to that used in the proof of Lemma 5.1. (One compensates for the fact that $\tilde{U}$ does not necessarily decrease exponentially in z_1 by using the decay estimate of Corollary 4.4.) As we shall observe later, it is not difficult to prove that these two integrals are also bounded by $C\tilde{z}_2^{-1}$ (uniformly in $\tilde{z}_1$) as $\tilde{z}_2 \to \infty$. A combination of these two facts yields that the right hand side in the formula (26) converges to zero as $|\tilde{z}| \to \infty$, which now implies that it is indeed a representation of $\tilde{U}$. For $a_0/2 < \tilde{z}$ we calculate

$$\begin{aligned}\frac{\partial}{\partial z_2}\tilde{U}(\tilde{z}) = \frac{a_0^2-1}{4\pi a_0^2}\int_{\{(s,t):\ 0<t<a_0/2\}\setminus\tilde{E}}\Bigg[\frac{2(\tilde{z}_2-t)}{|\tilde{z}_1-s|^2+|\tilde{z}_2-t|^2} \\ -\frac{2(\tilde{z}_2+t)}{|\tilde{z}_1-s|^2+|\tilde{z}_2+t|^2}\Bigg]\left(\frac{\partial}{\partial s}\right)^2\tilde{U}(s,t)\,dsdt \\ +\frac{1}{4\pi}\int_{\tilde{E}}\left[\frac{2(\tilde{z}_2-t)}{|\tilde{z}_1-s|^2+|\tilde{z}_2-t|^2}-\frac{2(\tilde{z}_2+t)}{|\tilde{z}_1-s|^2+|\tilde{z}_2+t|^2}\right]g(s,t)\,dsdt.\end{aligned}$$

Introducing $K(x,y) = \frac{2y}{x^2+y^2}$, we obtain

$$\begin{aligned}\frac{\partial}{\partial z_2}\tilde{U}(\tilde{z}) = -\frac{a_0^2-1}{4\pi a_0^2}\int_{\{(s,t)\ :\ 0<t<a_0/2\}\setminus\tilde{E}} 2tK_y(\tilde{z}_1-s,\tilde{z}_2+\theta)\left(\frac{\partial}{\partial s}\right)^2\tilde{U}(s,t)\,dsdt \\ -\frac{1}{4\pi}\int_{\tilde{E}}2tK_y(\tilde{z}_1-s,\tilde{z}_2+\theta)g(s,t)\,dsdt,\end{aligned}$$

with θ lying between $-a_0/2$ and $a_0/2$ (and depending on $\tilde{z}_1 - s$, $\tilde{z}_2$, and t). Since $|K_y(x,y)| = |\frac{2(x^2-y^2)}{(x^2+y^2)^2}| \le Cy^{-2}, 1 < y$, the known decay of $\left(\frac{\partial}{\partial s}\right)^2 \tilde{U}$ (Corollary 4.4) and the boundedness of g imply

$$\left|\frac{\partial}{\partial \tilde{z}_2}\tilde{U}(\tilde{z})\right| \le C\tilde{z}_2^{-2}, \quad 1 + a_0/2 < \tilde{z}_2. \tag{27}$$

An argument identical to that given just above (taking no derivative) would immediately yield that the two integrals in the right hand of the formula (26) are bounded by $C\tilde{z}_2^{-1}$. (We needed this fact earlier when we showed that (26) is indeed a representation of $\tilde{U}$.) Rewritten in terms of the z coordinates, (27) gives

$$\left|\frac{\partial}{\partial z_2}U(z)\right| \le Cz_2^{-2}, \quad 1/2 + c < z_2, \tag{28}$$

for some $0 < c$. Using the fact that

$$\left(\frac{\partial}{\partial z_2}A(z_2)\frac{\partial}{\partial z_2}U(z)\right) = -A(z_2)\left(\frac{\partial}{\partial z_1}\right)^2 U(z) = O(|z|^{-3}), \quad z \in \Phi(\Omega_+), \tag{29}$$

as asserted by Corollary 5.4, we are now able to prove the following lemma.

Lemma 6.1. *There exists a constant C, such that*

$$\left|\frac{\partial}{\partial z_2}U(z)\right| \le C|z|^{-2}$$

for $z \in \Phi(\Omega_+) \setminus \{z_2 = 1/2\}$.

Proof. From (28) it follows immediately that

$$\left|\frac{\partial}{\partial z_2}U(z)\right| \le Cz_2^{-2} \le C|z|^{-2} \quad \text{for } \max(1/2 + c, |z_1|) \le z_2. \tag{30}$$

For $0 < z_2$ and $\max(K, z_2) \le |z_1|$ (with K sufficiently large) we have that

$$\left(A\frac{\partial}{\partial z_2}U\right)(z_1,z_1) - \left(A\frac{\partial}{\partial z_2}U\right)(z_1,z_2) = \int_{z_2}^{z_1}\left(\frac{\partial}{\partial z_2}A\frac{\partial}{\partial z_2}U\right)(z_1,t)\,dt,$$

and therefore

$$\begin{aligned}\left|\left(A\frac{\partial}{\partial z_2}\right)U(z_1,z_2)\right| &\le C|z|^{-2} + C\int_{z_2}^{z_1}(|z_1|^2 + |t|^2)^{-3/2}\,dt\\ &\le C|z|^{-2} \quad \text{for } \max(K, z_2) \le |z_1|.\end{aligned} \tag{31}$$

Here we have used the estimates (29) and (30) (and the fact that we may select $K > 1/2 + c$) to derive the first inequality. Based on a combination of (30) and (31) we conclude that

$$\left|\frac{\partial}{\partial z_2}U(z)\right| \le C|z|^{-2} \quad \text{for } |z| > \sqrt{2}K, \ z \notin \{z_2 = 1/2\}.$$

The function A is discontinuous across $\{z_2 = 1/2\}$ and the derivative $\frac{\partial}{\partial z_2}U$ is not properly defined there; this is why we subtract the set $\{z_2 = 1/2\}$. The above estimate in combination with an elliptic regularity estimate (for $|z|$ small) immediately leads to the desired result. $\square$

Combining Corollary 5.4 and Lemma 6.1 we finally arrive at the following theorem.

THEOREM 6.2. *The solution, $u \in H^1(\Omega)$, to the boundary value problem* (1)–(2)*, with conductivity a given by* (3) *(and ϕ smooth) is in $W^{1,\infty}(\Omega)$ for any fixed $0 < a_0 < \infty$.*

Proof. We already know (cf. [4]) that $u \in C^{\beta}(\Omega)$ for some $\beta > 0$. From standard elliptic regularity results we also know that u is smooth, and therefore bounded, near $\partial\Omega$. It thus suffices to prove that $\nabla u \in L^{\infty}(\Omega)$. As already explained earlier (in section 2) we may restrict attention to u that are odd in the x_1-axis. For such u, it suffices to show that $|\nabla u| \leq C$ in $\Omega_+ \setminus \{x_1^2 + (x_2-1)^2 = 1\} = (\Omega \cap \{0 < x_2\}) \setminus \{x_1^2 + (x_2-1)^2 = 1\}$. The solution u has the form

$$u(x) = U \circ \Phi(x),$$

where U has been studied in the preceding three sections. We calculate

$$\nabla u(x) = D\Phi^t(x)(\nabla_z U)(\Phi(x)). \tag{32}$$

The matrix $D\Phi$ is given by

$$D\Phi = \begin{pmatrix} \dfrac{\partial z_1}{\partial x_1} & \dfrac{\partial z_1}{\partial x_2} \\ \dfrac{\partial z_2}{\partial x_1} & \dfrac{\partial z_2}{\partial x_2} \end{pmatrix},$$

and a simple computation yields

$$\left|\frac{\partial z_i}{\partial x_j}\right| = \left|\frac{\delta_{ij}(x_1^2+x_2^2) - 2x_i x_j}{(x_1^2+x_2^2)^2}\right| \leq C\frac{1}{x_1^2+x_2^2}, \qquad x \in \Omega_+. \tag{33}$$

At the same time, Corollary 5.4 and Lemma 6.1 give that

$$|\nabla_z U(\Phi(x))| \leq C|\Phi(x)|^{-2} = C(x_1^2+x_2^2) \quad \text{for } x \in \Omega_+ \setminus \{x_1^2 + (x_2-1)^2 = 1\}. \tag{34}$$

Combining (32), (33), and (34) we finally obtain

$$|\nabla u(x)| \leq C \qquad x \in \Omega_+ \setminus \{x_1^2 + (x_2-1)^2 = 1\},$$

as desired. $\square$

Remark. In the appendix we shall see that the case which formally corresponds to $a_0 = 0$ admits solutions that are discontinuous at the origin. Thus, it would not be reasonable to expect the solution u (given fixed boundary data) to have a gradient that is uniformly bounded, independently of a_0. The L^{∞} norm of $|\nabla u|$ may well become unbounded as a_0 approaches 0. By duality the same phenomenon may also occur as a_0 approaches ∞.

7. Appendix.

In this appendix we give a short review of what happens in the two cases that at least formally correspond to $a_0 = 0$ and $a_0 = \infty$. In both cases the relevant boundary value problems live in $\Omega \setminus \{\text{the fibers}\}$. They require that

$$\triangle u^0 = \triangle u^{\infty} = 0 \quad \text{in } \Omega \setminus \{\text{the fibers}\}, \tag{35}$$

with

$$\frac{\partial}{\partial n} u^0 = 0 \quad \text{on the boundaries of the fibers,} \tag{36}$$

and

$$u^\infty = \text{ constant on the boundary of each fiber,} \tag{37}$$

respectively. The constants in the boundary condition (37) are not arbitrary; they are those (or rather, it is that) for which the energy expression attains its smallest value. On the boundary $\partial\Omega$, the two solutions satisfy the common boundary condition

$$u^0 = u^\infty = \phi.$$

Given smooth boundaries, the solutions u^0 and u^∞ would be obtained as limits of the solution to (1), as a_0 tends to 0 and as a_0 tends to ∞, respectively. We suspect that the same holds true for boundaries with cusps as here, but we have not carried out the analysis. This is why we use the terminology "formally corresponding to $a_0 = 0$ and $a_0 = \infty$."

In the transformed variables $z = \Phi(x)$, with Φ as before, (35)–(37) become

$$\triangle U^0 = \triangle U^\infty = 0 \quad \text{in } \{z \in \Phi(\Omega), \; -1/2 < z_2 < 1/2\}, \tag{38}$$

with

$$\frac{\partial}{\partial z_2} U^0 = 0 \quad \text{on } z_2 = \pm 1/2 \tag{39}$$

and

$$U^\infty = c_\pm \quad \text{on } z_2 = \pm 1/2, \tag{40}$$

respectively. The common boundary condition on $\partial\Omega$ transforms into

$$U^0 = U^\infty = \phi \circ \Phi^{-1} \quad \text{on } \Phi(\partial\Omega).$$

For the moment we restrict attention to the boundary value problem for U^0. At the very end of this section, we return to make some remarks about the boundary value problem for U^∞. As mentioned previously, any solution to this boundary value problem may be written as a sum of two harmonic functions in $\{z \in \Phi(\Omega), -1/2 < z_2 < 1/2\}$, one which is even in the z_1-axis and one which is odd. These two functions have somewhat different behavior. We first consider the even function, which, when restricted to the interval $0 < z_2 < 1/2$, is a solution to

$$\begin{aligned} \triangle U^0 = \quad & \text{in } \{z \in \Phi(\Omega), \;\; 0 < z_2 < 1/2\}, \\ \frac{\partial}{\partial z_2} U^0 = 0 \quad & \text{on } \{z_2 = 1/2\} \text{ and on } \{z \in \Phi(\Omega), \;\; z_2 = 0\}, \\ U^0 = \phi \circ \Phi^{-1} \quad & \text{on } \{z \in \Phi(\partial\Omega), \;\; 0 < z_2\}. \end{aligned} \tag{41}$$

Separation of variables now immediately gives that U^0 must have the form

$$U^0(z_1, z_2) = \beta_0 + \sum_{n=1}^\infty \beta_n \cos(2n\pi z_2) e^{-2n\pi z_1} \quad \text{for } z_1 \text{ sufficiently positive,} \tag{42}$$

$$U^0(z_1, z_2) = \beta_0' + \sum_{n=1}^\infty \beta_n' \cos(2n\pi z_2) e^{2n\pi z_1} \quad \text{for } z_1 \text{ sufficiently negative.} \tag{43}$$

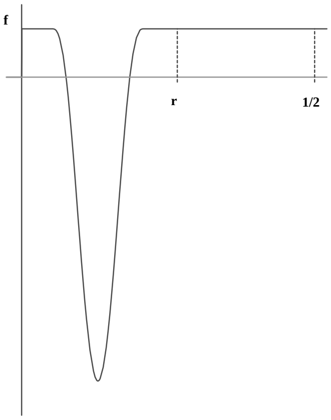


Fig. 10. *The function f.*

Conversely, any function, U^0, that is defined by (42) for $z_1 > 0$ and by (43) for $z_1 < 0$ is a solution to

$$\triangle U^0 = 0 \quad \text{in } \{z_1 \neq 0\}$$

with boundary conditions

$$\frac{\partial}{\partial z_2} U^0 = 0 \quad \text{on } \{z_2 = 0\} \text{ and on } \{z_2 = 1/2\}.$$

We shall now use this fact to construct a rather large class of solutions.

Select β_0 and β_0' arbitrarily, and let $f(z_2)$ denote any smooth, even, and periodic function with period 1, such that $f(z_2) = \beta_0' - \beta_0$ for $r < z_2 < 1/2$ $(0 < r)$ and such that $\int_0^{1/2} f(s)\, ds = 0$. A graph of such a function on the interval $(0, 1/2)$ is illustrated in Figure 10. The value of r is selected small enough, so that the line segment $\{z_1 = 0, \;\; 0 < z_2 < r\}$ lies inside $\Phi(\mathbb{R}^2 \setminus \Omega)$.

Let β_n, $n \geq 1$, be the cosine Fourier coefficients of the function $f/2$, i.e.,

$$2 \sum_{n=1}^{\infty} \beta_n \cos(2n\pi z_2) = f(z_2). \tag{44}$$

Since the integral of f is zero, the expansion does not contain any 0th order term. Since f is smooth, the β_n converge very fast to zero.

Let β'_n, $n \geq 1$, be given by

$$\beta'_n = -\beta_n, \tag{45}$$

and consider U^0 defined by (42) for $z_1 > 0$, and by (43) for $z_1 < 0$. Due to (44), (45), and the fact that $f(z_2) = \beta'_0 - \beta_0$ for $r < z_2 < 1/2$, we observe that U^0 is continuous across the line segment $\{z_1 = 0, \quad r < z_2 < 1/2\}$. The fact that $\beta'_n = -\beta_n$ ensures that $\frac{\partial}{\partial z_1} U^0$ is even in z_1 and thus automatically continuous across the line segment $\{z_1 = 0, r < z_2 < 1/2\}$. We conclude that this U^0 is indeed harmonic in all of $\Phi(\Omega) \cap \{0 < z_2 < 1/2\}$ and satisfies the boundary conditions

$$\frac{\partial}{\partial z_2} U^0 = 0 \quad \text{on } \{z_2 = 1/2\} \quad \text{and on } \{z \in \Phi(\Omega), \quad z_2 = 0\}.$$

The values of U^0 on $\Phi(\partial\Omega) \cap \{0 < z_2\}$ (the remainder of the boundary) naturally depend on f and correspond to a particular choice of ϕ.

Since β_0 and β'_0 were chosen arbitrarily we have $\beta'_0 \neq \beta_0$ in general. In the z coordinates β_0 and β'_0 are the limits of U^0 at $z_1 = +\infty$ and $z_1 = -\infty$, respectively. In the x-coordinates these are the limits of u^0 as we approach the origin through the cusp on the right and through the cusp on the left, respectively.

We have thus constructed a family of solutions which are discontinuous at the origin (and are even in the x_1-axis). They exhibit the behavior typicial of solutions when the data ϕ has an odd component with respect to the x_2-axis (and is even in the x_1-axis).

If the data ϕ is even in both the x_2- and the x_1-axis, then we must necessarily have $\beta'_n = \beta_n$ for all $0 \leq n$ (in the expansion (42) and (43)) and so u^0 has to be continuous at the origin. If the data ϕ is odd in the x_1-axis then it is very easy, again by separation of variables, to see that u^0 is continuous at the origin (its value is zero).

In all the cases considered above the function u^0 is C^∞ inside each of the cusps and all its derivatives (of order ≥ 1) vanish at the origin.

Let us now briefly return to the case $a_0 = \infty$. When ϕ is even in the x_1-axis, separation of variables readily gives that $U^\infty(z)$ must have the form

$$U^\infty(z_1, z_2) - c_0 = \sum_{n=1}^{\infty} \beta_n \cos((2n+1)\pi z_2) e^{-(2n+1)\pi z_1} \quad \text{for } z_1 \text{ sufficiently positive,}$$

$$U^\infty(z_1, z_2) - c_0 = \sum_{n=1}^{\infty} \beta'_n \cos((2n+1)\pi z_2) e^{(2n+1)\pi z_1} \quad \text{for } z_1 \text{ sufficiently negative,}$$

where c_0 is the common value attained on the fibers (there is just one constant value, due to the evenness of the solution). It follows immediately that u^∞ is continuous at 0 (in the x-coordinates), and that all its derivatives vanish at 0. When ϕ is odd in the x_1-axis, so are u^∞ and U^∞. Separation of variables thus yields

$$U^\infty(z_1, z_2) - 2c_0 z_2 = \sum_{n=1}^{\infty} \beta_n \sin(2n\pi z_2) e^{-2n\pi z_1} \quad \text{for } z_1 \text{ sufficiently positive,} \tag{46}$$

$$U^\infty(z_1, z_2) - 2c_0 z_2 = \sum_{n=1}^{\infty} \beta'_n \sin(2n\pi z_2) e^{2n\pi z_1} \quad \text{for } z_1 \text{ sufficiently negative,} \tag{47}$$

where c_0 is the value attained on the upper fiber. However, in this case the requirement that u^∞ be H^1 in the x-coordinates implies that the gradients ∇u^∞ and ∇U^∞ must be L^2 in the x- and in the z-coordinates, respectively. Thus, c_0 must be equal to 0. It follows, using the representation (46) and (47), that u^∞ is continuous at $x = 0$ (it has value 0), and that, similarly, all its derivatives vanish at $x = 0$.

The fact that all the solutions are C^∞ when regarded as functions in just each individual cusp would also follow from the analysis in [7].

As mentioned earlier, it would be very interesting to analyze the geometric setting when the fibers are close but not quite touching, say, the cross-sections are ϵ apart vertically. A few things can be said related to the calculations carried out above, as the distance ϵ tends to 0. When the boundary data ϕ has an odd component with respect to the x_2-axis and is otherwise even in the x_1-axis, then the singularity mentioned above for u^0 gives rise to a gradient (an x_1-derivative $\frac{\partial}{\partial x_1} u^0_\epsilon(0)$) which becomes unbounded as ϵ tends to zero. The solution u^∞ for the case of a ϕ, which is even in the x_2-axis but has an odd component with respect to the x_1-axis, is related to the previous solution by harmonic conjugation (rotation of the gradient by 90 degrees). We thus in general, in this case, also obtain a gradient (an x_2-derivative $\frac{\partial}{\partial x_2} u^\infty_\epsilon(0)$) which becomes unbounded as ϵ tends to zero. This in spite of the fact that there is no irregularity in the "limiting" solution when the fibers touch. The rate at which this gradient becomes unbounded has actually been calculated in [2], for a special solution corresponding to uniform antiplane shear (see also [9]). For this special solution, the rate turns out to be $\epsilon^{-1/2}$; we believe this is the generic rate for the above mentioned symmetries in the boundary data. It should also be mentioned that for two touching fibers and $0 < a_0 < \infty$, Budiansky and Carrier [2] calculate a finite value for the stress $\left(a \frac{\partial}{\partial x_2} u\right)(0)$ of the same special (antiplane shear) solution. This calculation relates $\left(a \frac{\partial}{\partial x_2} u\right)(0)$ to the "shear at infinity."

Acknowledgments. This work was initiated while Eric Bonnetier was visiting the Department of Mathematics at Rutgers University; it was completed while Michael Vogelius was visiting the Centre Emile Borel at the Institut Henri Poincaré. The hospitality of these two institutions is gratefully acknowledged. Both authors would like to thank Ivo Babuska of the University of Texas for stimulating discussions and communications that provoked their initial interest in this problem.

REFERENCES

[1] B. Andersson, *private communication*, Aeronautical Research Institute of Sweden, 1998.

[2] B. Budiansky and G. F. Carrier, *High shear stresses in stiff fiber composites*, Trans. ASME J. Appl. Mech., 51 (1984), pp. 733–735.

[3] E. DeGiorgi, *Sulla differenziabilità e l'analiticità delle estremali degli integrali multipli regolari*, Mem. Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur., 3 (1957), pp. 25–43.

[4] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1983.

[5] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[6] J. B. Keller, *Stresses in narrow regions*, Trans. ASME J. Appl. Mech., 60 (1993), pp. 1054–1056.

[7] V. A. Kozlov, V. G. Maz'ya, and J. Rossmann, *Elliptic Boundary Value Problems in Domains with Point Singularities*, Math. Surveys Monogr. 52, AMS, Providence, RI, 1989.

[8] Y.-Y. Li and M. Vogelius, *Gradient estimates for solutions to divergence form elliptic equations with discontinuous coefficients*, Arch. Rational Mech. Anal., to appear.

[9] X. Markenscoff, *Stress amplification in vanishingly small geometries*, Comput. Mech., 19 (1996), pp. 77–83.

[10] J. Nash, *Continuity of solutions of parabolic and elliptic equations*, Amer. J. Math., 80 (1958), pp. 931–954.

SIAM J. Math. Anal.
Vol. 31, No. 3, pp. 678–692

BEHAVIOR OF SOLUTIONS NEAR THE FLAT HATS OF STATIONARY SOLUTIONS FOR A DEGENERATE PARABOLIC EQUATION*

SHINGO TAKEUCHI†

Abstract. The behavior of solutions u of the degenerate parabolic equation $u_t = \lambda(|u_x|^{p-2}u_x)_x + |u|^{q-2}u(1-|u|^r)$, defined in $(0,1)\times(0,+\infty)$, is discussed. It is well known that there exists a stationary solution ϕ which has an open set Ω where it is identically ± 1. We call a graph $\{(x,\phi(x)); x\in\Omega\}$ flat hats. We investigate the behavior of $u(x,t)$ near $(x,t)\in\Omega\times[0,+\infty)$ where $|u(x,t)-\phi(x)|$ is very small. We will give a sufficient condition for initial data u_0 that the intersection points between the flat hats of ϕ and u never change as a function of t along $u(\cdot,t;u_0)$. Even if the condition failed, it is also proved that the changing area of the intersection points is uniformly bounded for t. Moreover we study stability properties for the positive stationary solution and the sign-changing stationary solutions.

Key words. degenerate parabolic equation, p-Laplace operator, flat hat, comparison theorem, intersection comparison

AMS subject classifications. 35K65, 35B40

PII. S003614109834257X

1. Problem and results.

In this paper we discuss the behavior of solutions of the following degenerate parabolic equation:

$$\text{(P)}\qquad \begin{cases} u_t = \lambda(|u_x|^{p-2}u_x)_x + f(u), & (x,t)\in(0,1)\times(0,+\infty),\\ u(0,t)=u(1,t)=0, & t\in(0,+\infty),\\ u(x,0)=u_0(x), & x\in(0,1), \end{cases}$$

where $p>2$ and λ is a positive parameter. The p-Laplace operator $\operatorname{div}(|\nabla u|^{p-2}\nabla u) = (|u_x|^{p-2}u_x)_x$ appears in the study of motions of non-Newtonian fluids in rheology (see, e.g., [2]). Many authors have studied (P) with $f\equiv 0$ or a source type $f(u)=|u|^{q-2}u$ (for example, [1, 3, 4, 7, 11, 12, 13, 14, 16]), and in particular, the latter case has been extensively investigated in view of a blow-up problem. For a source-absorption type $f(u)=|u|^{q-2}u(1-|u|^r)$, there are a few works by Guedda–Veron [8], Kamin–Veron [10], and Takeuchi–Yamada [15].

We will recall the work of [15] for $f(u)=|u|^{q-2}u(1-|u|^r)$ with $q\ge 2$ and $r>0$. For any initial datum $u_0\in L^2=L^2(0,1)$, there exists a unique global strong solution $u(\cdot;u_0)$ of (P), which belongs to $C([0,+\infty);L^2)\cap C((0,+\infty);W_0^{1,p})$ [15, Theorem 2.1]. We can also see that $u(t;u_0)$ approaches its ω-limit set $\omega(u_0)$ as $t\to+\infty$ and that $\omega(u_0)$ is included in E_λ, which is the set of all solutions of

$$\text{(SP)}\qquad \begin{cases} \lambda(|\phi_x|^{p-2}\phi_x)_x + f(\phi) = 0, & x\in(0,1),\\ \phi(0)=\phi(1)=0 \end{cases}$$

[15, Theorem 2.2], so that $u(t;u_0)$ converges to E_λ as $t\to+\infty$.

*Received by the editors July 28, 1998; accepted for publication (in revised form) May 7, 1999; published electronically March 6, 2000. This work was partially supported by JSPS Fellowships for Japanese Young Scientists.

http://www.siam.org/journals/sima/31-3/34257.html

†Department of Mathematics, Waseda University 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, Japan (stakeuchi@nyc.odn.ne.jp).

The structure of E_λ in the degenerate diffusion case $p > 2$ is very different from that in the linear diffusion case $p = 2$. For $p = 2$, all solutions of (SP) satisfy $|\phi(x)| < 1$ for $x \in [0, 1]$, E_λ is a discrete set for any $\lambda > 0$, and stability properties of all functions in E_λ are studied well (for example, see [9, pp. 118–128]). In case $p > 2$, if λ is sufficiently small, there exist solutions of (SP) for which the set $\{x \in [0, 1]; |\phi(x)| = 1\}$ is nonempty and E_λ consists of a discrete set and some continua ([15, Theorems 3.1–3.3], [8], and [10]). We will explain these facts more concretely. For $p > 2$, define

$$\Lambda_l := (l+1)\left\{\frac{\lambda(p-1)}{p}\right\}^{1/p}\int_0^1 (F(1)-F(\xi))^{-1/p}d\xi, \quad l = 0, 1, 2, \dots ,$$

where $F(\xi) = \int_0^\xi f(s)ds$. Then $\{\Lambda_l\}_{l=0}^\infty$ is well defined and is strictly increasing. For $\Lambda_{2l+1} < 1$, we define a family of open intervals $\{\Omega_k\}_{k=0}^l$ by

$$\Omega_k := \begin{cases} (\Lambda_0, \Lambda_0 + c_0), & k = 0, \\ (\Lambda_{2k} + \sum_{j=0}^{k-1} c_j, \Lambda_{2k} + \sum_{j=0}^{k} c_j), & k = 1, 2, \dots , l, \end{cases}$$

and a family of closed intervals $\{\Pi_k\}_{k=0}^{l+1}$ by

$$\Pi_k := \begin{cases} [0, \Lambda_0], & k = 0, \\ [\Lambda_{2k-2} + \sum_{j=0}^{k-1} c_j, \Lambda_{2k} + \sum_{j=0}^{k-1} c_j], & k = 1, 2, \dots , l, \\ [\Lambda_{2l} + \sum_{j=0}^{l} c_j, 1], & k = l + 1, \end{cases}$$

where $\{c_j\}_{j=0}^l$ is any finite sequence satisfying

$$c_j > 0 \quad \text{and} \quad \sum_{j=0}^{l} c_j = 1 - \Lambda_{2l+1}. \tag{1.1}$$

Note that $|\Omega_k| = c_k$ and

$$2|\Pi_0| = 2|\Pi_{l+1}| = |\Pi_k| = \Lambda_1, \qquad k = 1, 2, \dots , l,$$
$$[0, 1] = \bigcup_{k=0}^{l} \Omega_k \cup \bigcup_{k=0}^{l+1} \Pi_k \quad \text{(disjoint union)}.$$

When $\lambda > 0$ satisfies $\Lambda_{2l+1} < 1$ for some l, there exists a solution $\phi = \phi_l$ having exactly l-zero points in $(0, 1)$ with $\phi_x(0) > 0$ such that it is expressed as

$$\phi(x) = \begin{cases} (-1)^k, & x \in \Omega_k,\ k = 0, 1, 2, \dots , l, \\ g(x), & x \in \Pi_0, \\ (-1)^k g\left(x - \Lambda_{2k-1} - \sum_{j=0}^{k-1} c_j\right), & x \in \Pi_k,\ k = 1, 2, \dots , l, \\ (-1)^{l+1} g(x-1), & x \in \Pi_{l+1}, \end{cases} \tag{1.2}$$

where g is a continuously differentiable, strictly monotone increasing and odd function on $[-\Lambda_0, \Lambda_0]$ satisfying $g'(0) = \{pF(1)/\lambda(p-1)\}^{1/p}$, $g(\Lambda_0) = 1$, and $g'(\Lambda_0) = 0$. (Clearly, a stationary solution satisfying $\phi_x(0) < 0$ has a similar expression to (1.2).)

Since we can choose $\{c_j\}_{j=0}^{l}$ arbitrarily as long as it satisfies (1.1), such solutions generate a continuum in E_λ. In the expression (1.2), each Ω_k, $k = 0, 1, \dots, l$, is called a *flat core of* ϕ and we set $\Omega := \bigcup_{k=0}^{l} \Omega_k$. The graph of ϕ on each Ω_k is said to be a *flat hat of* ϕ, and the graph of ϕ on each Π_k is called a *layer of* ϕ. In [15], little information on the stability for such stationary solutions has been obtained though we have discussed stability properties of stationary solutions without flat hats. In this connection, the self-similar solutions with flat parts of degenerate parabolic equations are well known and play an important role for the blow-up solution. This is a so-called countable Q-set of profiles (cf. [1, 7]).

Our purpose is to study the behavior of nonstationary solution u of (P) when it stays in a neighborhood of a stationary solution with flat hats. Before stating our results, we will prepare a set of functions. Let $\Lambda_{2l+1} < 1$ and let ϕ be a solution of (SP) with l-zero points. For any compact subset I in Ω and any open neighborhood J of I in Ω, we define

$$U_0(I, J) := \{u_0 \in C([0,1]); u_0(x) = \phi(x) \text{ if } x \in I,\ u_0(x) \neq \phi(x) \text{ if } x \in J \setminus I\}.$$

For any (x,t) where solution u of (P) intersects flat hats of ϕ, the reaction effect for u disappears and there exists only the diffusion effect for u, whose coefficient is $\lambda(p-1)|u_x|^{p-2}$. When u_0 touches ϕ anywhere in its flat hats, we can expect that $u(t; u_0)$ keeps on touching ϕ there and that the touching area does not spread out. In fact, this is right if u_0, which touches the flat hats, is very close to ϕ in the following sense.

Theorem 1.1. *Let $\Lambda_{2l+1} < 1$ and let ϕ be a solution of* (SP) *with flat hats and l-zero points. Take any $\varepsilon \in (0, r)$, any connected compact subset $I := [a, b]$ in Ω, any open neighborhood J of I in Ω, and any open neighborhoods $N(a)$ and $N(b)$ in J. Then there exists $\delta_0 = \delta_0(\varepsilon, I, J, N(a), N(b)) > 0$ with the following property*:

> *For any $\delta \in (0, \delta_0)$ there exists an open neighborhood J_δ of I in J such that if $u_0 \in U_0(I, J)$, $\|u_0 - \phi\|_\infty < \delta$, and for $z = a$ and $z = b$*
>
> $$|u_0(x) - \phi(x)| \le \left(\frac{p-2}{p}\right)^{p/(p-2)} \left\{\frac{p(r-\varepsilon)}{2\lambda(p-1)}\right\}^{1/(p-2)} |x - z|^{p/(p-2)}, \quad x \in N(z),$$
>
> *then $u(\cdot, t; u_0) \in U_0(I, J_\delta)$ for all $t \in [0, +\infty)$ and $|J \setminus J_\delta| \to 0$ as $\delta \to 0$.*

In case u_0 crosses a flat hat transversely, the diffusion may cause their intersection points to change as a function of t along $u(t; u_0)$. The following result assures that the area on which the intersections may change is uniformly bounded for t.

Theorem 1.2. *Let $\Lambda_{2l+1} < 1$ and let ϕ be a solution of* (SP) *with flat hats and l-zero points. Take any point x_0 in Ω and any open neighborhood J of x_0 in Ω. Then there exists $\delta_0 = \delta_0(x_0, J) > 0$ with the following property*:

> *For any $\delta \in (0, \delta_0)$ there exist open sets P_δ and J_δ such that $x_0 \in P_\delta \subset \overline{P_\delta} \subset J_\delta \subset J$ and that, if $u_0 \in U_0(\{x_0\}, J)$ and $\|u_0 - \phi\|_\infty < \delta$, then*
>
> $$u(x, t; u_0) \neq \phi(x) \qquad \textit{for } (x,t) \in (J_\delta \setminus P_\delta) \times [0, +\infty),$$
>
> *and $|P_\delta|$, $|J \setminus J_\delta| \to 0$ as $\delta \to 0$.*

The proofs of Theorems 1.1 and 1.2 rely on the comparison theorem. Phase-plain analysis for (SP) will help us to find suitable comparison functions. Combining the proofs of these theorems, we can see the behavior of $u(\cdot; u_0)$ for more general u_0.

Theorems 1.1 and 1.2 give information about the motion of intersection points between u and the flat hats of ϕ. We can regard the analysis for intersections as one of *intersection comparison* (see [5, 6, 14]). When the asymptotic behavior is concerned, it is very important to study the stability of the positive stationary solution ϕ_0 with a flat hat. Our previous result [15] says only that ϕ_0 is attractive with respect to L^∞-norm. In this paper we can also prove that ϕ_0 is asymptotically stable in the following sense.

THEOREM 1.3. *Let ϕ_0 be the positive stationary solution with a flat hat. For any $\varepsilon > 0$, there exists $\delta > 0$ such that $\|u_0 - \phi_0\|_\infty < \delta$ implies $\|u(t; u_0) - \phi_0\|_\infty < \varepsilon$ for all $t \in [0, +\infty)$. Furthermore, $u(t; u_0) \to \phi_0$ in L^∞ as $t \to +\infty$.*

From these theorems for flat hats, we may assert that flat hats of stationary solutions give little disturbance to nonstationary solutions. The proof of Theorem 1.3 uses the following *weak comparison theorem*, which enables us to choose piecewise smooth functions as comparison functions. (The definitions of a *weak upper solution* and a *weak lower solution* are given in Definition 3.1.)

THEOREM 1.4. *Let u and v be a weak upper and a weak lower solution of* (P) *in $[0, T]$ for initial data u_0 and v_0, respectively. If $u_0(x) \geq v_0(x)$ almost everywhere (a.e.) $x \in (0, 1)$, then $u(x, t) \geq v(x, t)$ for a.e. $(x, t) \in (0, 1) \times (0, T)$.*

A weak comparison theorem for (P) with $f \equiv 0$ is given by DiBenedetto [3] and Kilpeläinen–Lindqvist [11]. Though results of this sort don't seem to be entirely new, we will make sure of its proof.

Owing to Theorem 1.4, we can also obtain information on stability properties of sign-changing stationary solutions ϕ_l ($l \geq 1$) with flat hats. In [15], we have shown that the sign-changing solutions with adjoining layers (i.e., $c_j = 0$ for some j in (1.1)) are all unstable in Lyapunov's sense. On the other hand, if we are concerned with the sign-changing solutions whose layers are separated by flat hats, no information has been derived for stability properties of each solution. In the present paper, we will prove that each ϕ_l is *conditionally stable*, i.e., stable for initial data that satisfy a certain condition for the shape near layers of ϕ_l.

The content of this paper is as follows. In section 2, constructing comparison functions with the aid of the phase-plain analysis, we prove Theorems 1.1 and 1.2. Moreover, a remark for more general u_0 is given. Section 3 is devoted to the proof of a weak comparison theorem. In sections 4 and 5, we show Theorem 1.3 and give some stability properties of sign-changing stationary solutions with flat hats by choosing suitable comparison functions.

2. Local behavior of solutions near flat hats. In this section we show Theorems 1.1 and 1.2 by the comparison theorem in [15] with use of suitable comparison functions and give a comment on the local behaviors of nonstationary solutions. First of all, we prepare the following lemma, which plays an important role in constructing comparison functions.

LEMMA 2.1. *For each $M > 0$, there exists a unique function $h^* = h^*_M$ satisfying the following properties*:

(i) *there exists a unique $x_M > 0$ such that $h^*(x_M) = M + 1$,*

(ii) *h^* is strictly monotone increasing in $(0, x_M)$,*

(iii) *h^* is a solution of the initial value problem*:

$$\begin{cases} \lambda(|h^*_x|^{p-2} h^*_x)_x + f(h^*) = 0, & x \in (0, x_M), \\ h^*(0) = 1, \quad h^*_x(0) = 0, \end{cases}$$

(iv) $h^*_{M_1}(x) = h^*_{M_2}(x)$ *for* $x \in (0, M_1)$, *where* $0 < M_1 \leq M_2$.
Furthermore, for any $\varepsilon \in (0, r)$ *there exists* $\theta = \theta(h^*) \in (0, x_M)$ *such that*

$$C_\varepsilon^- x^{p/(p-2)} \leq h^*(x) - 1 \leq C_\varepsilon^+ x^{p/(p-2)} \quad \textit{for } x \in [0, \theta), \tag{2.1}$$

where

$$C_\varepsilon^\pm = \left(\frac{p-2}{p}\right)^{p/(p-2)} \left\{\frac{p(r \pm \varepsilon)}{2\lambda(p-1)}\right\}^{1/(p-2)}.$$

Proof. The first half of the assertion is directly deduced from the phase plain analysis (see [15]). We note that x_M is given by

$$x_M = \left\{\frac{\lambda(p-1)}{p}\right\}^{1/p} \int_1^{1+M} (F(1) - F(\xi))^{-1/p} d\xi < +\infty$$

for $p > 2$. Moreover, it is easy to see that h^* satisfies

$$x = \left\{\frac{\lambda(p-1)}{p}\right\}^{1/p} \int_1^{h^*(x)} (F(1) - F(\xi))^{-1/p} d\xi \quad \text{for } x \in (0, x_M). \tag{2.2}$$

Observe that for any $\varepsilon \in (0, r)$ there exists $\eta > 0$ such that

$$\left(\frac{r-\varepsilon}{2}\right)(\xi - 1)^2 \leq F(1) - F(\xi) \leq \left(\frac{r+\varepsilon}{2}\right)(\xi - 1)^2 \quad \text{for } \xi \in (1, 1+\eta). \tag{2.3}$$

Thus taking small $\theta \in (0, x_M)$ such that $h^*(x) < 1 + \eta$ for $x \in (0, \theta)$, we have by (2.2) and (2.3)

$$\left(\frac{2}{r+\varepsilon}\right)^{1/p} \leq \frac{p-2}{p}\left\{\frac{p}{\lambda(p-1)}\right\}^{1/p} \frac{x}{(h^*(x)-1)^{1-2/p}} \leq \left(\frac{2}{r-\varepsilon}\right)^{1/p}$$

for $x \in (0, \theta)$. Therefore we obtain (2.1). □

Remark 2.1. Let g be a function appearing in (1.2). One can also show that for any $\varepsilon \in (0, r)$, g satisfies

$$C_\varepsilon^-(\Lambda_0 - x)^{p/(p-2)} \leq 1 - g(x) \leq C_\varepsilon^+(\Lambda_0 - x)^{p/(p-2)} \quad \text{for } x \in (\Lambda_0 - \theta, \Lambda_0)$$

with some $\theta = \theta(g) \in (0, \Lambda_0)$, where C_ε^- and C_ε^+ are the same constants as in Lemma 2.1.

Setting $h_*(x) := -h^*(-x)$ and $y_M := \Lambda_0 - g^{-1}(1 - M)$ for $M > 0$, we will show Theorems 1.1 and 1.2.

Proof of Theorem 1.1. We discuss only the case $\phi(x) \equiv 1$ on I since the case $\phi(x) \equiv -1$ on I is treated in the same way.

Let $I = [a, b]$, $J = (a_0, b_0)$, $N(a) \setminus (a, b) = (a_1, a]$, $N(b) \setminus (a, b) = [b, b_1)$ with $a_0 < a_1 < a \leq b < b_1 < b_0$ and let $\varepsilon \in (0, r)$ be any number. Take any $M > 0$ and fix it. Define

$$\phi^*(x; M, I) := \begin{cases} h^*(-x + a), & x \in [a - x_M, a], \\ 1, & x \in [a, b], \\ h^*(x - b), & x \in (b, b + x_M], \\ +\infty, & x \in [0, a - x_M) \cup (b + x_M, 1], \end{cases}$$

and

$$(2.4)\quad \phi_*(x;M,I) := \begin{cases} h_*(x-a+\Lambda_1), & x \in [a-\Lambda_1-x_M, a-\Lambda_1),\\ g(x-a+\Lambda_0), & x \in [a-\Lambda_1, a),\\ 1, & x \in [a,b],\\ -g(x-b-\Lambda_0), & x \in (b, b+\Lambda_1],\\ h_*(-x+b+\Lambda_1), & x \in (b+\Lambda_1, b+\Lambda_1+x_M],\\ -\infty, & x \in [0, a-\Lambda_1-x_M) \cup (b+\Lambda_1+x_M, 1], \end{cases}$$

where g is the function appearing in (1.2). By Lemma 2.1 and Remark 2.1 there exist $\theta(h^*) \in (0, x_M)$ and $\theta(g) \in (0, \Lambda_0)$ such that

$$(2.5)\qquad \begin{cases} C_\varepsilon^-(a-x)^{p/(p-2)} \le \phi^*(x;M,I) - 1, & x \in (a-\theta(h^*), a],\\ C_\varepsilon^-(x-b)^{p/(p-2)} \le \phi^*(x;M,I) - 1, & x \in [b, b+\theta(h^*)), \end{cases}$$

$$(2.6)\qquad \begin{cases} C_\varepsilon^-(a-x)^{p/(p-2)} \le 1 - \phi_*(x;M,I), & x \in (a-\theta(g), a],\\ C_\varepsilon^-(x-b)^{p/(p-2)} \le 1 - \phi_*(x;M,I), & x \in [b, b+\theta(g)). \end{cases}$$

Put $\theta := \min\{a-a_1, b_1-b, \theta(h^*), \theta(g)\}$. Fix $\delta_0 > 0$ such that $\delta_0 < \min\{h^*(\theta)-1, 1-g(\theta)\}$, $\delta_0 < \phi(x) - \phi_*(x;M,I)$ for $x \in [0,1] \setminus J$ and that $x_{\delta_0},\ y_{\delta_0} < \min\{a-a_0, b_0-b\}$. Note $\delta_0 \in (0, \min\{1, M\})$. We will show that δ_0 satisfies the assertion.

Let $\delta \in (0, \delta_0)$ and assume $u_0 \in U_0(I,J)$, $\|u_0 - \phi\|_\infty < \delta$ and

$$(2.7)\qquad \begin{cases} |u_0(x) - 1| \le C_\varepsilon^-(a-x)^{p/(p-2)}, & x \in (a_1, a],\\ |u_0(x) - 1| \le C_\varepsilon^-(x-b)^{p/(p-2)}, & x \in [b, b_1). \end{cases}$$

From (2.5), (2.6), and (2.7) we have

$$\phi_*(x;\delta_0,I) \le u_0(x) \le \phi^*(x;\delta_0,I), \quad x \in (a-\theta, b+\theta).$$

In addition, since h^* and g are strictly increasing functions, we see

$$(2.8)\qquad \begin{cases} u_0(x) \le \phi^*(x;M,I), & x \in (a-x_M, b+x_M),\\ u_0(x) \ge \phi_*(x;M,I), & x \in (a-\Lambda_1-x_M, b+\Lambda_1+x_M). \end{cases}$$

Here we have used the fact that $u_0(x) > \phi(x) - \delta > \phi(x) - \delta_0 > \phi_*(x;M,I)$ for $x \in [0,1] \setminus J$. If $\|u_0 - \phi\|_\infty < \delta$, then $-1-\delta < u_0(x) < 1+\delta$ for $x \in [0,1]$; therefore $u \equiv 1+\delta$ (resp., $u \equiv -1-\delta$) is an upper solution (resp., a lower solution) for (P). Thus the comparison theorem [15, Theorem 2.3] yields that $|u(x,t;u_0)| \le 1+\delta$ for $(x,t) \in [0,1] \times [0,+\infty)$; in particular,

$$(2.9)\qquad \begin{cases} u(x,t;u_0) \le 1+\delta \le \phi^*(x;M,I) = 1+M\\ \qquad\qquad\qquad \text{at } x = a-x_M \text{ and } x = b+x_M,\\ u(x,t;u_0) \ge -1-\delta \ge \phi_*(x;M,I) = -1-M\\ \qquad\qquad\qquad \text{at } x = a-\Lambda_1-x_M \text{ and } x = b+\Lambda_1+x_M \end{cases}$$

for all $t \in [0, +\infty)$. Therefore, by virtue of (2.8) and (2.9), the comparison theorem gives

(2.10)
$$\begin{cases} u(x,t;u_0) \leq \phi^*(x;M,I), & (x,t) \in [a - x_M, b + x_M] \times [0,+\infty), \\ u(x,t;u_0) \geq \phi_*(x;M,I), & (x,t) \in [a - \Lambda_1 - x_M, b + \Lambda_1 + x_M] \times [0,+\infty); \end{cases}$$

therefore we conclude

$$u(x,t;u_0) = 1 = \phi(x) \text{ for all } (x,t) \in I \times [0,+\infty). \tag{2.11}$$

It remains to show that there exists an open neighborhood J_δ of I in J such that $u(x,t;u_0) \neq \phi(x)$ for $x \in J_\delta \setminus I$. The set $J \setminus I$ consists of two connected sets $[a_0, a)$ and $(b, b_0]$. Without loss of generality we can assume that

$$\begin{aligned} u_0(x) &< \phi(x) \quad \text{if } x \in [a_0, a), \\ u_0(x) &> \phi(x) \quad \text{if } x \in (b, b_0]. \end{aligned} \tag{2.12}$$

We will show that there exist $a_\delta \in (0, a-a_0)$ and $b_\delta \in (0, b_0-b)$, which are independent of u_0, such that

$$u(x,t;u_0) < \phi(x) \quad \text{if } x \in [a_0 + a_\delta, a), \tag{2.13}$$
$$u(x,t;u_0) > \phi(x) \quad \text{if } x \in (b, b_0 - b_\delta] \tag{2.14}$$

for all $t \in [0,+\infty)$ and that $a_\delta,\ b_\delta \to 0$ as $\delta \to 0$. Take any a_δ satisfying $x_\delta < a_\delta < x_{\delta_0}$ $(< \min\{a - a_0, b_0 - b\})$. Then we have

$$u_0(x) \leq \phi^*(x; M, [a_0 + a_\delta, a]), \quad x \in [a_0 + a_\delta - x_M, a + x_M].$$

From (2.12), for any $\xi \in (0, a - a_0 - a_\delta)$ there exists $\eta \in (0, h^*(a_\delta) - 1 - \delta)$ such that

$$u_0(x) < \phi^*(x; M, [a_0 + a_\delta, a - \xi]) - \eta, \quad x \in [a_0 + a_\delta - x_M, a - \xi + x_M]. \tag{2.15}$$

Define

$$u^*(x,t) := \phi^*(x; M, [a_0 + a_\delta, a - \xi]) - \eta e^{-Rt}$$

for $(x,t) \in [a_0 + a_\delta - x_M, a - \xi + x_M] \times [0,+\infty)$, where R is a positive number such that $R > -f'(1+M)$. Then

$$u(x,t;u_0) \leq 1 + \delta < u^*(x,t) \quad \text{at } x = a_0 + a_\delta - x_M \text{ and } x = a - \xi + x_M$$

for all $t \in [0,+\infty)$ and by an easy calculation

$$\begin{aligned} &u^*_t - \lambda(|u^*_x|^{p-2}u^*_x)_x - f(u^*) \\ &\quad = \eta R e^{-Rt} + f(\phi^*) - f(\phi^* - \eta e^{-Rt}) \\ &\quad \geq \eta e^{-Rt}(R + f'(1+M)) \\ &\quad > 0; \end{aligned}$$

therefore u^* is an upper solution of (P). Therefore, it follows from the comparison theorem that

$$u(x,t;u_0) \leq u^*(x,t) < \phi^*(x; M, [a_0 + a_\delta, a - \xi]) = \phi(x)$$

for $(x,t) \in [a_0 + a_\delta, a - \xi] \times [0, +\infty)$. Since $\xi \in (0, a - a_0 - a_\delta)$ is arbitrary, we can conclude

$$u(x,t;u_0) < \phi(x) \text{ for } (x,t) \in [a_0 + a_\delta, a) \times [0, +\infty).$$

We can choose a_δ such that $a_\delta \to 0$ as $\delta \to 0$; therefore (2.13) holds true.

The proof of (2.14) is similar to that of (2.13) if we use the following comparison function u_* instead of u^*. Let b_δ be any number satisfying $y_\delta < b_\delta < y_{\delta_0}$ and $\phi_*(x; M, [b, b_0 - b_\delta]) < \phi(x) - \delta$ for $x \in [0,1] \setminus J$. For any $\xi \in (0, b_0 - b_\delta - b)$ there exists $\eta \in (0, 1 - g(\Lambda_0 - b_\delta) - \delta)$ such that

$$u_0(x) > \phi_*(x; M, [b + \xi, b_0 - b_\delta]) + \eta, \quad x \in [b + \xi - \Lambda_1 - x_M, b_0 - b_\delta + \Lambda_1 + x_M]. \tag{2.16}$$

We have only to define

$$u_*(x,t) := \phi_*(x; M, [b + \xi, b_0 - b_\delta]) + \eta e^{-Rt}$$

for $(x,t) \in [b + \xi - \Lambda_1 - x_M, b_0 - b_\delta + \Lambda_1 + x_M] \times [0, +\infty)$, where $R > -f'(1 + M)$. Thus the proof is accomplished. □

Proof of Theorem 1.2. We give the proof in the case $\phi(x_0) = 1$. Let $J = (a_0, b_0)$ with $a_0 < x_0 < b_0$ and fix $\delta_0 > 0$ satisfying $2x_{\delta_0}$, $2y_{\delta_0} < \min\{x_0 - a_0, b_0 - x_0\}$. Without loss of generality we may assume that

$$u_0(x) < \phi(x) \quad \text{if } x \in [a_0, x_0), \tag{2.17}$$

$$u_0(x) > \phi(x) \quad \text{if } x \in (x_0, b_0]. \tag{2.18}$$

Then we can show the assertion by following the arguments used in the latter half of the proof of Theorem 1.1 with $a = b = x_0$. However, we should take $\xi = a_\delta$ in case (2.17) (resp., $\xi = b_\delta$ in case (2.18)). For such ξ, it is possible to take η satisfying (2.15) because $2a_\delta < 2x_{\delta_0} < \min\{x_0 - a_0, b_0 - x_0\}$ (resp., (2.16) because $2b_\delta < 2y_{\delta_0} < \min\{x_0 - a_0, b_0 - x_0\}$). □

We will give a remark associated with Theorems 1.1 and 1.2. In the proofs of these theorems, $\phi^*(\cdot; M, I)$ and $\phi_*(\cdot; M, I)$ are essential. Let $u_0 \in C([0,1])$. If only u_0 satisfies (2.8) for some $I = [a,b]$ $(\subset \Omega)$ and $M > \max\{\|u_0\|_\infty - 1, 0\}$, then $u(\cdot, t; u_0)$ satisfies (2.10) and hence (2.11). Similarly if u_0 satisfies

$$\begin{gathered} u_0(x) < \phi^*(x; M, I), \quad x \in (a - x_M, b + x_M) \\ (\text{resp., } u_0(x) > \phi_*(x; M, I), \quad x \in (a - \Lambda_1 - x_M, b + \Lambda_1 + x_M)) \end{gathered}$$

for some $I = [a,b]$ $(\subset \Omega)$ and the same M above, then in a similar way as (2.15) (resp., (2.16)) we obtain

$$\begin{gathered} u(x,t;u_0) < \phi(x) \quad \text{for } (x,t) \in I \times [0, +\infty) \\ (\text{resp., } u(x,t;u_0) > \phi(x) \quad \text{for } (x,t) \in I \times [0, +\infty)). \end{gathered}$$

Therefore when u_0 is given, it is possible to see the local behavior of $u(x,t;u_0)$ in $\Omega \times [0, +\infty)$ as long as we can take such comparison functions ϕ^* and ϕ_*. We can weaken the assumption $\|u_0 - \phi\|_\infty < \delta$ as u_0 lies in a neighborhood of ϕ on a local domain. Indeed, to know the behavior of intersection points between the flat hats of ϕ and u, we have only to take ϕ^* and ϕ_* as above.

3. Weak comparison theorem. In this section we show a weak comparison theorem (Theorem 1.4). The definition of a weak upper or lower solution of (P) is given by the following.

DEFINITION 3.1. *If $u \in C([0,T];L^2) \cap L^p(0,T;W^{1,p}) \cap L^{q+r}(0,T;L^{q+r})$ satisfies the following inequalities, then it is called a weak upper solution of* (P) *in* $[0,T]$*:*

$$\tag{3.1} \begin{cases} \displaystyle\int_0^1 u\varphi(x,t)dx + \int_0^t \int_0^1 (-u\varphi_t + |u_x|^{p-2}u_x\varphi_x)(x,\tau)dxd\tau \\ \qquad \displaystyle\geq \int_0^1 u_0\varphi(x,0)dx + \int_0^t \int_0^1 f(u)\varphi(x,\tau)dxd\tau \qquad \textit{for all } t \in [0,T], \\ u(0,t) \geq 0, \quad u(1,t) \geq 0 \qquad \textit{for a.e. } t \in (0,T) \end{cases}$$

for all functions $\varphi \in W^{1,2}(0,T;L^2) \cap L^p(0,T;W_0^{1,p}) \cap L^\infty([0,1]\times[0,T])$ satisfying $\varphi \geq 0$. A function u is called a weak lower solution if (3.1) *holds true with "$\geq$" replaced by "$\leq$."*

Remark 3.1. (i) In Definition 3.1, $u \in L^\alpha(0,T;L^\alpha)$ with $\alpha = q+r$; therefore $f(u)$ belongs to $L^{\alpha/(\alpha-1)}(0,T;L^{\alpha/(\alpha-1)}) \subset L^1(0,T;L^1)$. Therefore the second term of the right-hand side of (3.1) converges. (ii) The strong solution of (P) becomes a weak upper and a weak lower solution (see [15]).

Remark 3.2. Let $0 = a_0 < a_1 < \cdots < a_n < a_{n+1} = 1$ and let $\{u_i : [a_i, a_{i+1}] \to \mathbb{R}\}$ be a family of functions such that

$$u_i(a_{i+1}) = u_{i+1}(a_{i+1}), \qquad i = 0, 1, \dots, n-1,$$

and

$$\tag{3.2} \begin{cases} u_{i,t} \geq \lambda(|u_{i,x}|^{p-2}u_{i,x})_x + f(u_i), & (x,t) \in (a_i, a_{i+1}) \times (0,T), \\ u_0(0,t) \geq 0,\ u_n(1,t) \geq 0, & t \in (0,T), \\ u_{i,x}(a_{i+1}-0,t) \geq u_{i+1,x}(a_{i+1}+0,t), & t \in (0,T). \end{cases}$$

If u is defined by $u = u_i$ for $x \in [a_i, a_{i+1}]$, then it is easily seen that u is a weak upper solution of (P). If u satisfies (3.2) with "$\geq$" replaced by "$\leq$," then u becomes a weak lower solution of (P).

The following lemma plays an important role in proving Theorem 1.4.

LEMMA 3.1. *Let u be a weak upper solution of* (P) *in* $[0,T]$ *and take any $h \in (0,T)$. Then u satisfies the inequalities*

$$\tag{3.3} \int_0^1 ([u]_{h,t}\varphi + [|u_x|^{p-2}u_x]_h\varphi_x - [f(u)]_h\varphi)(x,t)dx \geq 0$$

for all $t \in (0, T-h)$ and for all $\varphi \in W_0^{1,p}$ satisfying $\varphi \geq 0$, where $[\phi]_h$ means the Steklov average of ϕ, i.e.,

$$[\phi]_h(x,t) := \begin{cases} \displaystyle\frac{1}{h}\int_t^{t+h} \phi(x,\tau)d\tau, & t \in (0, T-h], \\ 0, & t \in (T-h, T). \end{cases}$$

For a weak lower solution v, (3.3) *holds true with "$\geq$" replaced by "$\leq$."*

Proof. Take any $t \in (0, T-h)$ and $\varphi \in W_0^{1,p}$ satisfying $\varphi \geq 0$. Define the following cut-off function:

$$\eta^n(\tau) := \begin{cases} 0, & \tau \in [0, t-1/n), \\ \displaystyle\int_{t-\tau}^{1/n} \rho^n(s)ds, & \tau \in [t-1/n, T], \end{cases}$$

where $\rho^n(s) := n\rho(ns)$ and $\rho \in C_0^\infty(-\infty, +\infty)$ is a function satisfying that $\rho(s) \geq 0$, $\rho(s) = 0$ if $|s| \geq 1$, and $\int_{-\infty}^{+\infty} \rho(s)ds = 1$. Note that $\eta^n(\tau) = 1$ for all $\tau \in [t+1/n, T]$.

We set

$$\varphi^n(x,\tau) := \varphi(x)\eta^n(\tau)$$

for $(x,\tau) \in [0,1] \times [0,T]$ and $n \geq 1$. Since $\varphi^n \in C^1([0,T]; W_0^{1,p})$, we have (3.1) with φ and t replaced by φ^n and $t+h$, respectively:

$$\begin{aligned} &\int_0^1 u(x,t+h)\varphi(x)\eta^n(t+h)dx \\ &\quad + \int_0^{t+h}\int_0^1 \{-u(x,\tau)\varphi(x)\eta_t^n(\tau) + |u_x(x,\tau)|^{p-2}u_x(x,\tau)\varphi_x(x)\eta^n(\tau)\}dxd\tau \\ &\quad \geq \int_0^{t+h}\int_0^1 f(u(x,\tau))\varphi(x)\eta^n(\tau)dxd\tau. \end{aligned} \tag{3.4}$$

In (3.4) it is easy to see

$$\int_0^1 u(x,t+h)\varphi(x)\eta^n(t+h)dx \longrightarrow \int_0^1 u(x,t+h)\varphi(x)dx$$

and

$$\begin{aligned} &\int_0^{t+h}\int_0^1 f(u(x,\tau))\varphi(x)\eta^n(\tau)dxd\tau \\ &\quad = \int_{t-1/n}^{t+1/n} \eta^n \int_0^1 f(u)\varphi dxd\tau + \int_{t+1/n}^{t+h}\int_0^1 f(u)\varphi dxd\tau \\ &\quad \longrightarrow \int_t^{t+h}\int_0^1 f(u)\varphi(x)dxd\tau \end{aligned}$$

as $n \to +\infty$. The remaining term is expressed as

$$\begin{aligned} &\int_{t-1/n}^{t+1/n}\int_0^1 (-u\varphi\eta_t^n + |u_x|^{p-2}u_x\varphi_x\eta^n)dxd\tau + \int_{t+1/n}^{t+h}\int_0^1 |u_x|^{p-2}u_x\varphi_x dxd\tau \\ &\quad = -\int_{t-1/n}^{t+1/n} \rho^n(t-\tau)\int_0^1 u\varphi dxd\tau + \int_{t-1/n}^{t+1/n} \eta^n \int_0^1 |u_x|^{p-2}u_x\varphi_x dxd\tau \\ &\qquad + \int_{t+1/n}^{t+h}\int_0^1 |u_x|^{p-2}u_x\varphi_x dxd\tau. \end{aligned}$$

One can show that the right-hand side of the above equality converges to

$$-\int_0^1 u(x,t)\varphi(x)dx + \int_t^{t+h}\int_0^1 |u_x|^{p-2}u_x\varphi_x dxd\tau \quad \text{as } n \to +\infty.$$

Therefore, letting $n \to +\infty$ in (3.4) we obtain

$$\begin{aligned}\int_0^1 u(x,t+h)\varphi(x)dx - \int_0^1 u(x,t)\varphi(x)dx + \int_t^{t+h}\int_0^1 |u_x|^{p-2}u_x\varphi_x dx d\tau \\ \geq \int_t^{t+h}\int_0^1 f(u)\varphi(x)dxd\tau.\end{aligned} \tag{3.5}$$

Dividing (3.5) by h and recalling the definition of Steklov average we get

$$\int_0^1 [u]_{h,t}\varphi(x)dx + \int_0^1 [|u_x|^{p-2}u_x]_h\varphi_x dx \geq \int_0^1 [f(u)]_h\varphi(x)dx. \qquad \square$$

Proof of Theorem 1.4. Let u and v be a weak upper and a weak lower solution of (P). For every $\tau \in (0,t)$ such that $u(\tau),\ v(\tau) \in W^{1,p}$ and $h \in (0, T-\tau)$, Lemma 3.1 yields

$$\begin{aligned}\int_0^1 \{([v]_{h,t} - [u]_{h,t})(\tau)\varphi \\ + ([|v_x|^{p-2}v_x]_h - [|u_x|^{p-2}u_x]_h)(\tau)\varphi_x - ([f(v)]_h - [f(u)]_h)(\tau)\varphi\}dx \leq 0\end{aligned}$$

for any $\varphi \in W_0^{1,p}$. Putting $\varphi(x) = ([v(x,\tau) - u(x,\tau)]_h)_+ \in W_0^{1,p}$, where $w_+(x) := \max\{w(x), 0\}$, we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_0^1 ([v-u]_h)_+^2(x,\tau)dx + \int_0^1 [|v_x|^{p-2}v_x - |u_x|^{p-2}u_x]_h([v-u]_h)_{+,x}dx \\ \leq \int_0^1 [f(v)-f(u)]_h([v-u]_h)_+dx.\end{aligned}$$

Integrating the above inequality over $(0,t)$ and letting $h \to 0$ (see [3, Chapter I, Lemma 3.2]), we see that the second term of the left-hand side becomes nonnegative because of the monotonicity of the p-Laplace operator. Therefore

$$\begin{aligned}\frac{1}{2}\int_0^1 (v-u)_+^2(x,t)dx - \frac{1}{2}\int_0^1 (v_0-u_0)_+^2(x)dx \\ \leq \int_0^t\int_0^1 (f(v)-f(u))(v-u)_+dxd\tau \\ \leq C\int_0^t\int_0^1 (v-u)_+^2(x,\tau)dxd\tau,\end{aligned}$$

where $C = \sup\{f'(s); s \in \mathbb{R}\} < +\infty$. By Gronwall's inequality,

$$\|(v-u)_+(t)\| \leq \|(v_0-u_0)_+\|e^{Ct}.$$

By the assumption, the right-hand side of this inequality vanishes; therefore the assertion follows. $\square$

Remark 3.3. In sections 4 and 5, we will apply Theorem 1.4 by choosing continuous functions (in $[0,1]\times[0,T]$) as weak upper and lower solutions. Therefore, the comparison holds for all $t \in [0,T]$.

4. Stability of ϕ_0. In this section we prove Theorem 1.3 by using Theorem 1.4.

Proof of Theorem 1.3. Let ϕ_0 be the positive solution of (SP) with a flat hat. Since the attractivity has been proved in [15], it suffices to show the stability of ϕ_0.

Let any $\varepsilon > 0$ be fixed. First we will show the stability of ϕ_0 from above. For any $\eta > 0$, we define

$$\phi_\eta^*(x) := \min\{-\phi_*(x+\eta;\eta,\{-\Lambda_0\}), 1+\eta, -\phi_*(x-\eta;\eta,\{1+\Lambda_0\})\}$$

for $x \in [0,1]$, where ϕ_* is the function appearing in (2.4). If $\eta > 0$ is sufficiently small, then

$$\phi_0(x) < \phi_\eta^*(x) < \phi_0(x) + \varepsilon, \quad x \in [0,1].$$

Setting $u^*(x,t) := \phi_\eta^*(x)$ for $(x,t) \in [0,1] \times [0,+\infty)$, we see from Remark 3.2 that u^* is a weak upper solution. Therefore, taking $\delta > 0$ such that $\phi_0(x) + \delta < \phi_\eta^*(x)$ for $x \in [0,1]$, we conclude from Theorem 1.4 that for any u_0 satisfying $u_0(x) < \phi_0(x) + \delta$

$$u(x,t;u_0) \le u^*(x,t) = \phi_\eta^*(x) < \phi_0(x) + \varepsilon \text{ for } (x,t) \in [0,1] \times [0,+\infty). \tag{4.1}$$

Next we will prove the stability of ϕ_0 from below. Consider the following initial value problem:

$$\text{(IP)} \qquad \begin{cases} \lambda(|\phi_x|^{p-2}\phi_x)_x + f(\phi) = 0, \\ \phi(0) = 0, \quad \phi_x(0) = \alpha, \end{cases}$$

where α is a positive parameter. It is known in [15] that the solution $\phi(\cdot;\alpha)$ of (IP) is a continuously differentiable and strictly monotone increasing function on $[0, X(\alpha)]$, where $X(\alpha)$ satisfies $\phi_x(X(\alpha);\alpha) = 0$. Note that the function g, which generates the layers of ϕ_0, is identical with $\phi(\cdot;\alpha_0)$ on $[0, X(\alpha_0)] = [0, \Lambda_0]$, where $\alpha_0 := \{pF(1)/\lambda(p-1)\}^{1/p}$. Let $\xi > 0$ be any sufficiently small number and fix it. We can see in [15] that $X(\alpha_0 - \xi) < X(\alpha_0)$ and that $\phi(x;\alpha_0 - \xi) < \phi_0(x)$ for $x \in (0, X(\alpha_0 - \xi))$. Putting

$$\tilde{\phi}(x;\alpha_0-\xi) := \begin{cases} -\infty, & x \in (-\infty, -X(\alpha_0-\xi)), \\ -\phi(-x;\alpha_0-\xi), & x \in [-X(\alpha_0-\xi), 0), \\ \phi(x;\alpha_0-\xi), & x \in [0, X(\alpha_0-\xi)], \\ +\infty, & x \in (X(\alpha_0-\xi), +\infty), \end{cases}$$

we define

$$\phi_*^\eta(x) := \min\{\tilde{\phi}(x-\eta;\alpha_0-\xi), \tilde{\phi}(X(\alpha_0-\xi);\alpha_0-\xi), -\tilde{\phi}(x-1+\eta;\alpha_0-\xi)\}$$

for $x \in [0,1]$. If $\eta \in (0, X(\alpha_0) - X(\alpha_0 - \xi))$ is sufficiently small, one can show

$$\phi_0(x) - \varepsilon < \phi_*^\eta(x) < \phi_0(x), \quad x \in [0,1].$$

Observe that $u_*(x,t) := \phi_*^\eta(x)$ is a weak lower solution of (P) in $[0,1] \times [0,+\infty)$. Therefore, taking $\delta > 0$ such that $\phi_*^\eta(x) < \phi_0(x) - \delta$ for $x \in [0,1]$, one can see from Theorem 1.4 that for any u_0 satisfying $\phi_0(x) - \delta < u_0(x)$

$$\phi_0(x) - \varepsilon < \phi_*^\eta(x) = u_*(x,t) \le u(x,t;u_0) \text{ for } (x,t) \in [0,1] \times [0,+\infty). \tag{4.2}$$

Finally, combining (4.1) and (4.2), we see that $\|u_0 - \phi_0\|_\infty < \delta$ implies $\|u(t;u_0) - \phi_0\|_\infty < \varepsilon$ for all $t \in [0,+\infty)$. □

5. Conditional stability of ϕ_l. In this section we will discuss a conditional stability of each sign-changing solution ϕ_l of (SP) with flat hats by using the comparison functions used in the previous sections.

We will discuss the case $l = 1$. Recall that ϕ_1 is expressed as

$$\phi_1(x) = \begin{cases} g(x), & x \in \Pi_0 = [0, \Lambda_0), \\ 1, & x \in \Omega_0 = [\Lambda_0, \Lambda_0 + c_0], \\ -g(x - \Lambda_1 - c_0), & x \in \Pi_1 = (\Lambda_0 + c_0, \Lambda_2 + c_0), \\ -1, & x \in \Omega_1 = [\Lambda_2 + c_0, \Lambda_2 + c_0 + c_1], \\ g(x-1), & x \in \Pi_2 = (\Lambda_2 + c_0 + c_1, 1], \end{cases}$$

where c_0 and c_1 are any positive numbers satisfying $c_0 + c_1 = 1 - \Lambda_3 > 0$. We introduce the following two functions:

$$\overline{\phi^*_\delta}(x) := \min\{-\phi_*(x+\delta; \delta, \{-\Lambda_0\}), 1+\delta, -\phi_*(x-\delta; \delta, \{\Lambda_2 + c_0\})\}, \quad x \in [0,1],$$

and

$$\overline{\phi_*^\delta}(x) := \begin{cases} +\infty, & x \in [0, \Lambda_1 + c_0 + \delta), \\ \max\{-\tilde{\phi}(x - \Lambda_1 - c_0 - \delta; \alpha_0 - \xi), & \\ \qquad -\tilde{\phi}(X(\alpha_0 - \xi); \alpha_0 - \xi), \tilde{\phi}(x - 1 + \delta; \alpha_0 - \xi)\}, & x \in [\Lambda_1 + c_0 + \delta, 1]. \end{cases}$$

We define

$$\Phi^*_\delta(x) := \min\{\overline{\phi^*_\delta}(x), \overline{\phi_*^\delta}(x)\}, \quad x \in [0,1].$$

Similarly, making use of the two functions

$$\underline{\phi^*_\delta}(x) := \begin{cases} \min\{\tilde{\phi}(x-\delta; \alpha_0 - \xi), \tilde{\phi}(X(\alpha_0 - \xi); \alpha_0 - \xi), & \\ \qquad -\tilde{\phi}(x - \Lambda_1 - c_0 + \delta; \alpha_0 - \xi)\}, & x \in [0, \Lambda_1 + c_0 - \delta], \\ -\infty, & x \in (\Lambda_1 + c_0 - \delta, 1], \end{cases}$$

and

$$\underline{\phi_*^\delta}(x) := \max\{\phi_*(x+\delta; \{\Lambda_0 + c_0\}), -1-\delta, \phi_*(x-\delta; \{1 + \Lambda_0\})\}, \quad x \in [0,1],$$

we define

$$\Phi_*^\delta(x) := \max\{\underline{\phi^*_\delta}(x), \underline{\phi_*^\delta}(x)\}, \quad x \in [0,1].$$

Then ϕ_1 is sandwiched between Φ^*_δ and Φ_*^δ. More precisely, it holds that

$$\phi_1(\Lambda_0 + c_0 - \delta) = \Phi_*^\delta(\Lambda_0 + c_0 - \delta), \tag{5.1}$$

$$\phi_1(\Lambda_2 + c_0 + \delta) = \Phi^*_\delta(\Lambda_2 + c_0 + \delta) \tag{5.2}$$

and that $\Phi_*^\delta(x) < \phi_1(x) < \Phi^*_\delta(x)$ for all $x \in [0,1] \setminus \{\Lambda_0 + c_0 - \delta, \Lambda_2 + c_0 + \delta\}$. Furthermore, we can observe that Φ^*_δ and Φ_*^δ are a weak upper and a weak lower solution, respectively. Therefore, for any u_0 whose graph lies in the shaded portion of Figure 5.1, the solution $u(x, t; u_0)$ keeps on staying within the portion; therefore ϕ_1 is conditionally stable for such initial data. We should note that Φ^*_δ and Φ_*^δ have "bumps" at $x = \Lambda_0 + c_0 - \delta$ and $\Lambda_2 + c_0 + \delta$ as seen in (5.1) and (5.2). These bumps make the shaded portion narrow and play a role making $u(t; u_0)$ cling to the layer of ϕ_1.

To see the conditional stability of ϕ_l for general $l \geq 1$, it is sufficient to form bumps near each layer as is exhibited in Figure 5.2 (the case $l = 2$).

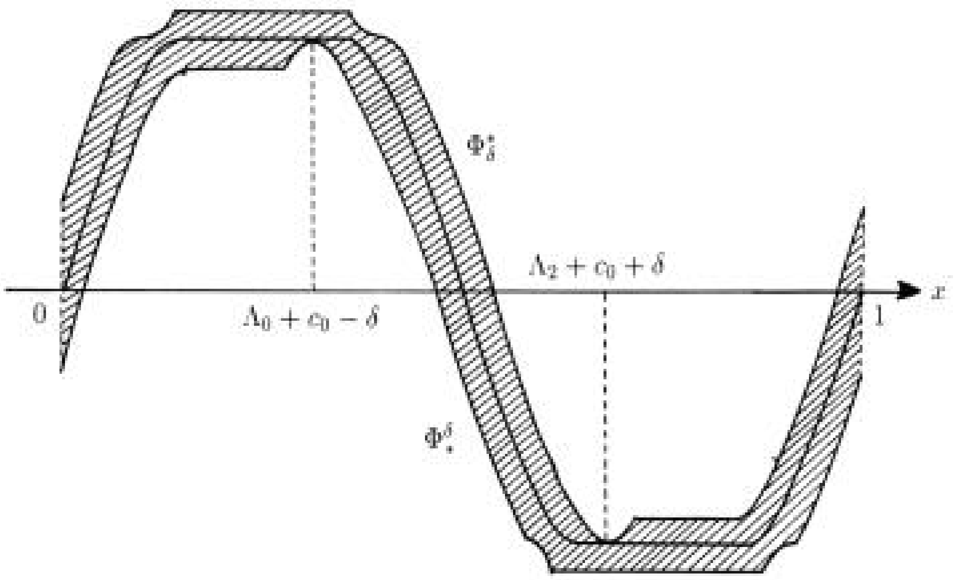


FIG. 5.1. *Domain of stability for ϕ_1.*

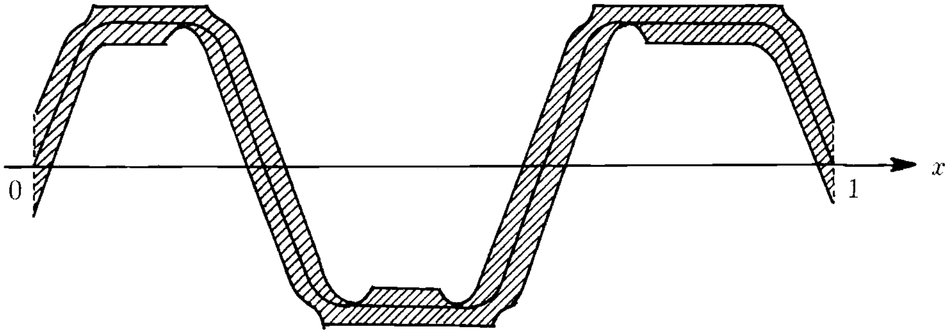


FIG. 5.2. *Domain of stability for ϕ_2.*

Acknowledgments. The author wishes to express his hearty thanks to Professor Yoshio Yamada for many useful suggestions. He also thanks the referees for some comments related to blow-up problems.

REFERENCES

[1] C. J. Budd and V. A. Galaktionov, *Stability and spectra of blow-up in problems with quasilinear gradient diffusivity*, Proc. Roy. Soc. London Ser. A, 454 (1998), pp. 2371–2407.

[2] J. I. Diaz, *Nonlinear Partial Differential Equations and Free Boundaries. Vol.* I *Elliptic Equations*, Res. Notes Math. 106, Pitman, Boston, London, 1985.

[3] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[4] A. Fujii and M. Ohta, *Asymptotic behavior of blowup solutions of a parabolic equation with the p-Laplacian*, Publ. Res. Inst. Math. Sci., 32 (1996), pp. 503–515.

[5] V. A. Galaktionov and J. L. Vazquez, *Extinction for a quasilinear heat equation with absorption* I. *Technique of intersection comparison*, Comm. Partial Differential Equations, 19 (1994), pp. 1075–1106.

[6] V. A. Galaktionov and J. L. Vazquez, *Geometric properties of the solutions of one-dimensional nonlinear parabolic equations*, Math. Ann., 303 (1995), pp. 741–769.

[7] V. A. Galaktionov, S. P. Kurdyumov, S. A. Posashkov, and A. A. Samarskii, *A nonlinear elliptic problem with a complex spectrum of solutions*, U.S.S.R. Comput. Math and Math. Phys., 26 (1986), pp. 48–54.

[8] M. Guedda and L. Veron, *Bifurcation phenomena associated to the p-Laplace operator*, Trans. Amer. Math. Soc., 310 (1988), pp. 419–431.

[9] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin/Heidelberg, 1981.

[10] S. Kamin and L. Veron, *Flat core properties associated to the p-Laplace operator*, Proc. Amer. Math. Soc., 118 (1993), pp. 1079–1085.

[11] T. Kilpeläinen and P. Lindqvist, *On the Dirichlet boundary value problem for a degenerate parabolic equation*, SIAM J. Math. Anal., 27 (1996), pp. 661–683.

[12] S. Sakaguchi, *The number of peaks of nonnegative solutions to some nonlinear degenerate parabolic equations*, J. Math. Anal. Appl., 203 (1996), pp. 78–103.

[13] S. Sakaguchi, *When are the spatial level surfaces of solutions of diffusion equations invariant with respect to the time variable?*, J. Anal. Math., 78 (1999), pp. 219–243.

[14] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-up in Quasilinear Parabolic Equations*, Nauka, Moscow, 1987 (in Russian); Walter de Gruyter, Berlin, 1995 (in English).

[15] S. Takeuchi and Y. Yamada, *Asymptotic properties of a reaction-diffusion equation with degenerate p-Laplacian*, Nonlinear Anal., to appear.

[16] M. Tsutsumi, *Existence and nonexistence of global solutions for nonlinear parabolic equations*, Publ. Res. Inst. Math. Sci., 8 (1972), pp. 211–229.

SIAM J. MATH. ANAL.
Vol. 31, No. 3, pp. 693–699

AN INEQUALITY INVOLVING THE GENERALIZED HYPERGEOMETRIC FUNCTION AND THE ARC LENGTH OF AN ELLIPSE[*]

ROGER W. BARNARD[†], KENT PEARCE[†], AND KENDALL C. RICHARDS[‡]

Abstract. In this paper we verify a conjecture of M. Vuorinen that the Muir approximation is a lower approximation to the arc length of an ellipse. Vuorinen conjectured that $f(x) = {}_2F_1(\frac{1}{2}, -\frac{1}{2}; 1; x) - [(1 + (1 - x)^{3/4})/2]^{2/3}$ is positive for $x \in (0, 1)$. The authors prove a much stronger result which says that the Maclaurin coefficients of f are nonnegative. As a key lemma, we show that ${}_3F_2(-n, a, b; 1 + a + b, 1 + \epsilon - n; 1) > 0$ when $0 < ab/(1 + a + b) < \epsilon < 1$ for all positive integers n.

Key words. hypergeometric, approximations, elliptical arc length

AMS subject classifications. 33C, 41A

PII. S0036141098341575

1. Introduction.

Let a and b be the semiaxes of an ellipse with eccentricity $e = \sqrt{a^2 - b^2}/a$. Let $L(a, b)$ denote the arc length of the ellipse. Without loss of generality we can take one of the semiaxes, say a, to be 1. Legendre's complete elliptic integral of the second kind can be defined by

$$E(r) = \int_0^{\pi/2} \sqrt{1 - r^2 \sin^2 t}\, dt.$$

Elliptic integrals are so named because of their connection with $L(a, b)$. In turn, these are related to Gauss's hypergeometric functions, ${}_2F_1$, defined by

$${}_2F_1(a_1, a_2; b_1; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n (a_2)_n}{(b_1)_n n!} z^n$$

with the Appell (or Pochhammer) symbol $(a)_n = a(a+1)\cdots(a+n-1)$ for $n \geq 1$ and $(a)_0 = 1, a \neq 0$. We shall need the generalized hypergeometric function, ${}_pF_q$, defined by

$${}_pF_q(a_1, a_2, \ldots, a_p; b_1, b_2, \ldots, b_q; z) = \sum_{n=0}^{\infty} \frac{(a_1)_n \cdot (a_2)_n \cdots (a_p)_n}{(b_1)_n \cdot (b_2)_n \cdots (b_q)_n} \frac{z^n}{n!}$$

(see [12, p. 73]). It was noted by Maclaurin in 1742 (see [2]) that

$$L(1, b) = 4E(e) = 2\pi {}_2F_1(\tfrac{1}{2}, -\tfrac{1}{2}; 1; e^2).$$

There are various references, books, and articles, which discuss the relationships between elliptic integrals and hypergeometric functions (see [3], [7]) and their role in

[*] Received by the editors July 7, 1998; accepted for publication (in revised form) December 17, 1998; published electronically March 6, 2000.

http://www.siam.org/journals/sima/31-3/34157.html

[†] Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409 (barnard@math.ttu.edu, pearce@math.ttu.edu).

[‡] Department of Mathematics, Southwestern University, Georgetown, TX 78626 (richards@southwestern.edu).

applications to physics (see [11], [9]) and in geometric function theory (see [10], [3]). From antiquity several more easily computable approximations to $L(a,b)$ have been suggested. The Almkvist–Berndt survey article [2] has an extensive discussion of these approximations. These approximations and their historical and recent connections to the approximations of π can be found in the Borweins' book [6]. An excellent source for all of the above ideas is the Anderson–Vamanamurthy–Vuorinen book *Conformal Invariants, Inequalities, and Quasiconformal Mappings* [3].

In 1883, it was proposed by Muir (see [2]) that $L(1,b)$ could be simply approximated by $2\pi[(1+b^{3/2})/2]^{2/3}$. A close numerical examination of the error in this approximation lead M. Vuorinen to pose Problem 5.6 in [13]. This was announced at several international conferences. Letting $x = 1-b^2$, he asked whether the Muir approximation

$$g(x) = \left(\frac{1+(1-x)^{3/4}}{2}\right)^{2/3}$$

is a lower approximation for the value given by the hypergeometric function

$$h(x) = {}_2F_1\left(\tfrac{1}{2}, -\tfrac{1}{2}; 1; x\right),$$

that is, whether

$$h(x) - g(x) \geq 0 \text{ for all } x \in (0,1).$$

We shall prove the following much stronger result.

THEOREM 1.1. *Let $g(x) = \sum_{n=0}^{\infty} a_n x^n$ and $h(x) = \sum_{n=0}^{\infty} A_n x^n$. Then,*

$$a_k \leq A_k \text{ for all } k = 0,1,2,\ldots,n,\ldots. \tag{1.1}$$

In particular, the function $f(x) \equiv [h(x)-g(x)]/x^4$ is convex and increasing from $(0,1]$ onto $(\alpha,\beta]$, where $\alpha = 2^{-14} = 0.000061\cdots$ and $\beta = (2/\pi) - 2^{-2/3} = 0.006659\cdots$.

Remarks. The ideas and techniques used to prove Lemma 2.1 and Theorem 1.1 will be used in [5] to determine surprising hierarchical relationships among the 13 historical approximations to $L(a,b)$ discussed in [2]. These approximations range over four centuries from Kepler's in 1642 to Almkvist's in 1985 and include two from Ramanujan.

2. Proof of main results. The proof of Theorem 1.1 requires the following lemma.

LEMMA 2.1. *Suppose $a,b>0$. Then, for any ϵ satisfying $\frac{ab}{1+a+b} < \epsilon < 1$,*

$${}_3F_2(-n,a,b;1+a+b,1+\epsilon-n;1) > 0 \quad \textit{for all integers } n \geq 1.$$

For the reader's convenience, we include the following classical identities.

Identity 1 (see [1, p. 558, eq. (15.2.24)]). If $|z|<1$, then

$$(c-b-1)\cdot {}_2F_1(a,b;c;z) = (c-1)\cdot {}_2F_1(a,b;c-1;z) - b\cdot\ {}_2F_1(a,b+1;c;z).$$

Identity 2 (see [12, p. 60, Thm. 21]). If $|z|<1$, then

$${}_2F_1(a,b;c;z) = (1-z)^{c-a-b}\cdot {}_2F_1(c-a,c-b;c;z).$$

Identity 3 (see [8, p. 59, eq. (3.1.1)]). If $F = {}_3F_2$, then

$$F(-n,a,b;c,d;1) = \frac{(d-b)_n}{(d)_n} F(-n,c-a,b;c,1+b-d-n;1).$$

Identity 4 (see [12, p. 82, eq. (14)]). If $F = {}_3F_2$ and $|z| < 1$, then

$$\begin{aligned}(a_1 - a_2) \cdot F(&a_1,a_2,a_3;b_1,b_2;z) \\ &= a_1 \cdot F(a_1+1,a_2,a_3;b_1,b_2;z) - a_2 \cdot F(a_1,a_2+1,a_3;b_1,b_2;z).\end{aligned}$$

Proof of Lemma 2.1. Using an idea suggested in [4], we let $F = {}_3F_2$ and consider the generating function

$$f(r) = \sum_{n=0}^{\infty} \frac{-(-\epsilon)_n}{n!} F(-n,a,b;1+a+b,1+\epsilon-n;1)r^n = \sum_{n=0}^{\infty} c_n r^n,$$

where $|r| < 1$. Note that $-(-\epsilon)_n > 0$ for $0 < \epsilon < 1$ and for all $n \geq 1$. Thus we seek to verify that $c_n > 0$ for all $n \geq 1$.

In this direction, we have

$$\begin{aligned}
f(r) &= \sum_{n=0}^{\infty} \frac{-(-\epsilon)_n}{n!} \sum_{k=0}^{n} \frac{(-n)_k (a)_k (b)_k}{(a+b+1)_k (1+\epsilon-n)_k k!} r^n \\
&= \sum_{n=0}^{\infty} \frac{-(-\epsilon)_n}{(1)_n} \sum_{k=0}^{n} \frac{\frac{(-1)^k (1)_n}{(1)_{n-k}} (a)_k (b)_k}{(a+b+1)_k \frac{(-1)^k (-\epsilon)_n}{(-\epsilon)_{n-k}} k!} r^n \\
&\qquad \left\{\text{using } (\alpha)_{n-k} = \tfrac{(-1)^k (\alpha)_n}{(1-\alpha-n)_k} \text{ and } (1)_n = n!\right\} \\
&= -\sum_{n=0}^{\infty} \sum_{k=0}^{n} \left(\frac{(a)_k (b)_k}{(a+b+1)_k k!} r^k \right) \left(\frac{(-\epsilon)_{n-k}}{(n-k)!} r^{n-k} \right) \\
&= -\sum_{n=0}^{\infty} \left(\frac{(-\epsilon)_n}{(n)!} r^n \right) \sum_{k=0}^{\infty} \left(\frac{(a)_k (b)_k}{(a+b+1)_k k!} r^k \right) \qquad \text{(see [12, p. 57, eq. (2)])} \\
&= -(1-r)^{\epsilon}\ {}_2F_1(a,b;a+b+1;r).
\end{aligned}$$

Differentiating, we have

$$\begin{aligned} f'(r) = \epsilon(1-r)^{\epsilon-1}\ &{}_2F_1(a,b;a+b+1;r) \\ - \frac{ab(1-r)^{\epsilon}}{(a+b+1)}\ &{}_2F_1(a+1,b+1;a+b+2;r). \end{aligned} \tag{2.1}$$

An application of Identity 1 followed by Identity 2 to ${}_2F_1(a+1,b+1;a+b+2;r)$ yields

$$\begin{aligned}
&\frac{ab(1-r)^{\epsilon}}{(a+b+1)} {}_2F_1(a+1,b+1;a+b+2;r) \\
&\quad = \frac{b(1-r)^{\epsilon}}{(a+b+1)} [(a+b+1) \cdot {}_2F_1(a+1,b+1;a+b+1;r) \\
&\quad - (b+1) \cdot {}_2F_1(a+1,b+2;a+b+2;r)] \\
&\quad = (1-r)^{\epsilon-1} \left[b \cdot {}_2F_1(a,b;a+b+1;r) - \frac{b(b+1)}{(a+b+1)} \cdot {}_2F_1(a,b+1;a+b+2;r) \right].
\end{aligned}$$

Thus (2.1) becomes

$$\begin{aligned}
f'(r) &= (1-r)^{\epsilon-1}\Bigg[(\epsilon-b)\cdot {}_2F_1(a,b;a+b+1;r) \\
&\quad + \frac{b(b+1)}{(a+b+1)}\cdot {}_2F_1(a,b+1;a+b+2;r)\Bigg] \\
&= (1-r)^{\epsilon-1}\sum_{n=0}^{\infty}\frac{(a)_n}{n!}\left[\frac{(b)_n(\epsilon-b)}{(a+b+1)_n}+\frac{b(b+1)(b+1)_n}{(a+b+1)(a+b+2)_n}\right]r^n \\
&= (1-r)^{\epsilon-1}\sum_{n=0}^{\infty}\frac{(a)_n}{n!}\left[\frac{(b)_n(\epsilon-b)}{(a+b+1)_n}+\frac{(b+1)(b)_n(b+n)}{(a+b+1)_n(a+b+1+n)}\right]r^n \qquad (2.2) \\
&= (1-r)^{\epsilon-1}\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(a+b+1)_n(a+b+1+n)n!} \\
&\quad \times\left[(\epsilon-b)(a+b+1+n)+(b+1)(b+n)\right]r^n \\
&= (1-r)^{\epsilon-1}\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(a+b+1)_n(a+b+1+n)n!} \qquad (2.3) \\
&\quad \times\left[\epsilon(a+b+1+n)+n-ab\right]r^n,
\end{aligned}$$

where (2.2) makes use of $\alpha(\alpha+1)_n=(\alpha)_n(\alpha+n)$. If $\frac{ab}{a+b+1}<\epsilon<1$, then the expression in (2.3) is the product of two series with all positive Maclaurin series coefficients. Hence f' has all positive Maclaurin series coefficients which is equivalent to the desired result. □

Corollary 2.2. *Let $T_n = {}_3F_2\left(-n,\frac{3}{2},\frac{1}{2};2,\frac{5}{4}-n;1\right)$. Then, for all integers $n\geq 8$,*

$$T_{n+1}>T_n>0.$$

Proof. Let $F={}_3F_2$ and $B_n=\left(\frac{3}{4}-n\right)_n/\left(\frac{5}{4}-n\right)_n$. Using Identity 3, we have that

$$T_n=B_nF\left(-n,\tfrac{1}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right).$$

Direct calculation reveals that $T_9>T_8>0>T_7>\cdots>T_2=T_1$. Now suppose that $T_n>T_{n-1}>0$ for some $n\geq 9$ and note that $B_{n+1}/B_n=\left(n+\frac{1}{4}\right)/\left(n-\frac{1}{4}\right)$. Then,

$$\begin{aligned}
T_{n+1} &= B_{n+1}F\left(-n-1,\tfrac{1}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right) \\
&= \frac{B_{n+1}}{\left(n+\frac{3}{2}\right)}\left[(n+1)F\left(-n,\tfrac{1}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right)+\tfrac{1}{2}F\left(-n-1,\tfrac{3}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right)\right] \qquad (2.4) \\
&= \frac{B_{n+1}(n+1)}{B_n\left(n+\frac{3}{2}\right)}T_n+\frac{B_{n+1}}{2\left(n+\frac{3}{2}\right)}F\left(-n-1,\tfrac{3}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right) \\
&= \frac{\left(n+\frac{1}{4}\right)(n+1)}{\left(n-\frac{1}{4}\right)\left(n+\frac{3}{2}\right)}T_n+\frac{B_{n+1}}{2\left(n+\frac{3}{2}\right)}F\left(-n-1,\tfrac{3}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right) \\
&> T_n+\frac{B_{n+1}}{2\left(n+\frac{3}{2}\right)}F\left(-n-1,\tfrac{3}{2},\tfrac{1}{2};2,\tfrac{1}{4};1\right),
\end{aligned}$$

where (2.4) follows from Identity 4, and the inequality holds because $\frac{(n+1/4)(n+1)}{(n-1/4)(n+3/2)} > 1$ and $T_n > 0$. Since $B_{n+1} < 0$, we shall have that $T_{n+1} > T_n > 0$ provided we show that $F(-n-1, \frac{3}{2}, \frac{1}{2}; 2, \frac{1}{4}; 1) < 0$. To this end, we again apply Identity 3 to observe that

$$F\left(-n-1, \tfrac{3}{2}, \tfrac{1}{2}; 2, \tfrac{1}{4}; 1\right) = \frac{\left(-\frac{1}{4}\right)_{n+1}}{\left(\frac{1}{4}\right)_{n+1}} F\left(-n-1, \tfrac{1}{2}, \tfrac{1}{2}; 2, \tfrac{1}{4}-n; 1\right).$$

Since

$$\frac{\left(-\frac{1}{4}\right)_{n+1}}{\left(\frac{1}{4}\right)_{n+1}} < 0,$$

we need to show that

$$F\left(-n-1, \tfrac{1}{2}, \tfrac{1}{2}; 2, \tfrac{1}{4}-n; 1\right) > 0.$$

Letting $m = n+1$, $a = b = \frac{1}{2}$, and $\epsilon = \frac{1}{4}$, it follows from Lemma 2.1 that

$$F\left(-n-1, \tfrac{1}{2}, \tfrac{1}{2}; 2, \tfrac{1}{4}-n; 1\right) = F(-m, a, b; a+b+1, 1+\epsilon-m; 1) > 0.$$

Hence $T_{n+1} > T_n > 0$ for all integers $n \geq 8$ by induction. □

Proof of Theorem 1.1. Clearly,

$$A_n = \frac{\left(\frac{1}{2}\right)_n \left(-\frac{1}{2}\right)_n}{n!n!}.$$

Computing the logarithmic derivative of g we have

$$\frac{g'(x)}{g(x)} = -\frac{1}{2}\left(\frac{(1-x)^{-\frac{1}{4}}}{1+(1-x)^{\frac{3}{4}}}\right),$$

which implies

$$\left(\sum_{n=0}^{\infty}(n+1)a_{n+1}x^n\right)\left((1-x)^{\frac{1}{4}}+1-x\right) = -\frac{1}{2}\sum_{n=0}^{\infty}a_n x^n. \tag{2.5}$$

The coefficients of x^n of the left-hand side of (2.5) are obtained from the Cauchy product of the two terms. Solving for a_{n+1} yields (by extracting the nth and $(n-1)$st terms from the Cauchy product)

$$a_{n+1} = \frac{1}{2(n+1)}\left[\left(\frac{5}{4}n-\frac{1}{2}\right)a_n - \sum_{k=0}^{n-2}(k+1)a_{k+1}\frac{\left(-\frac{1}{4}\right)_{n-k}}{(n-k)!}\right]. \tag{2.6}$$

We now verify (1.1) using an inductive argument. Clearly, the coefficients of the terms a_k in (2.6) are nonnegative. Computation gives: $a_0 = A_0 = 1,\ a_1 = A_1 = -1/4,\ a_2 = A_2 = -3/64,\ a_3 = A_3 = -5/2^8,\ a_4 = -11/2^{10}$ and $A_4 = -175/2^{14}$. Suppose that the inequality in (1.1) holds for $4 \leq k \leq n$. From (2.6) we have

$$\text{(2.7) } a_{n+1} \leq \frac{1}{2(n+1)}\left[\left(\frac{5}{4}n-\frac{1}{2}\right)\frac{\left(\frac{1}{2}\right)_n\left(-\frac{1}{2}\right)_n}{n!n!} - \sum_{k=0}^{n-2}(k+1)\frac{\left(\frac{1}{2}\right)_{k+1}\left(-\frac{1}{2}\right)_{k+1}}{(k+1)!(k+1)!}\frac{\left(-\frac{1}{4}\right)_{n-k}}{(n-k)!}\right].$$

We need to show that the right-hand side of (2.7) is less than or equal to $A_{n+1} = \frac{\left(\frac{1}{2}\right)_{n+1}\left(-\frac{1}{2}\right)_{n+1}}{(n+1)!(n+1)!}$, that is,

$$\begin{aligned}\left(\frac{5}{4}n-\frac{1}{2}\right)\frac{\left(\frac{1}{2}\right)_n\left(-\frac{1}{2}\right)_n}{n!n!} &- 2(n+1)\frac{\left(\frac{1}{2}\right)_{n+1}\left(-\frac{1}{2}\right)_{n+1}}{(n+1)!(n+1)!} \\ &\le \sum_{k=0}^{n-2}(k+1)\frac{\left(\frac{1}{2}\right)_{k+1}\left(-\frac{1}{2}\right)_{k+1}}{(k+1)!(k+1)!}\frac{\left(-\frac{1}{4}\right)_{n-k}}{(n-k)!}.\end{aligned} \tag{2.8}$$

After adding the $(n-1)$st and nth terms of the right-hand side of (2.8) to inequality (2.8) and then simplifying, we use $(a)_{k+1} = (a+k)(a)_k$, $(a)_{n-k} = (-1)^k(a)_k/(1-a-n)_k$, the fact that $(-n)_k = 0$ for $k \geq n+1$, and the definition of ${}_3F_2$ to obtain

$$\frac{\left(\frac{1}{2}\right)_n\left(-\frac{1}{2}\right)_n}{n!n!}\cdot\frac{(2n-1)}{4(n+1)} \le -\left(\frac{1}{4}\right)\left(-\frac{1}{4}\right)_n\frac{{}_3F_2\left(-n,\frac{1}{2},\frac{3}{2};2,\frac{5}{4}-n;1\right)}{n!},$$

or equivalently

$${}_3F_2\left(-n,\tfrac{1}{2},\tfrac{3}{2};2,\tfrac{5}{4}-n;1\right) \geq \frac{\left(\frac{1}{2}\right)_n^2}{\left(-\frac{1}{4}\right)_n(n+1)!}. \tag{2.9}$$

Clearly, the right-hand side of (2.9) is negative for all $n \geq 1$. Inequality (2.9) can be explicitly verified for $0 \leq n \leq 7$. For $n \geq 8$, inequality (2.9) follows from Corollary 2.2. Thus, the inequality in (1.1) also holds for $k = n+1$. Hence, by induction (1.1) holds for all $k \in \mathbb{N} \cup \{0\}$.

Finally, the convexity and monotonicity of f are clear. By l'Hôpital's rule, $f(0^+) = A_4 - a_4 = 1/2^{14} = 1/16384$, while the value of $f(1)$ is clear. □

Acknowledgments. The authors would like to thank the referees for their many helpful suggestions which helped to considerably improve the paper.

REFERENCES

[1] M. Abramowitz and I. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] G. Almkvist and B. Berndt, *Gauss, Landen, Ramanujan, the arithmetic-geometric mean, ellipses, pi, and the Ladies Diary,* Amer. Math. Monthly, 95 (1988), pp. 585–608.

[3] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Conformal Invariants, Inequalities, and Quasiconformal Mappings*, John Wiley, New York, 1997.

[4] R. Askey, G. Gasper, and M. Ismail, *A positive sum from summability theory,* J. Approx. Theory, 13 (1975), pp. 413–420.

[5] R. W. Barnard, K. Pearce, and K. C. Richards, *A monotonicity property involving ${}_3F_2$ and comparisons of classical approximations of elliptical arc length,* SIAM J. Math. Anal., to appear.

[6] J. M. Borwein and P. B. Borwein, *Pi and the AGM—A Study in Analytic Number Theory and Computational Complexity*, John Wiley, New York, 1987.

[7] B. C. Carlson, *Special Functions of Applied Mathematics*, Academic Press, New York, 1977.

[8] G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge University Press, Cambridge, 1990.

[9] A. G. Greenhill, *The Applications of Elliptic Functions*, Dover, New York, 1954.

[10] P. Henrici, *Applied and Computational Complex Analysis,* Vol. I*,* John Wiley, New York, 1974.

[11] D. F. Lawden, *Ellipitic Functions and Applications*, Appl. Math. Sci. 80, Springer-Verlag, New York, 1989.

[12] E. Rainville, *Special Functions*, Macmillan, New York, 1960.

[13] M. Vuorinen, *Hypergeometric functions in geometric function theory,* in Proceedings of the Special Functions and Differential Equations, K. R. Srinivasa, R. Jagannthan, and G. Van der Jeugy, eds., Allied Publishers, New Delhi, 1998.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 701–725

THE POLES OF THE RESOLVENT FOR THE EXTERIOR NEUMANN PROBLEM OF ANISOTROPIC ELASTICITY*

MISHIO KAWASHITA† AND GEN NAKAMURA‡

Abstract. The poles of the resolvent and the asymptotic behavior of the local energy for the exterior Neumann problem of elastic wave equations are considered. For the most general class of anisotropic elastic media, the existence of the poles approaching the real axis is proved if the Rayleigh surface waves exist at least locally. The rate of their convergence to the real axis is estimated. Some results which show that the local energy hardly escapes from any neighborhood of the boundary are also presented. These results are considered as an influence of the existence of the Rayleigh surface waves.

The local existence condition of the Rayleigh surface waves is given in terms of the surface impedance tensor, which is essentially equal to the principal part of the Neumann operator in the elliptic region. Unlike isotropic elastic media, the Rayleigh surface waves exist only locally for anisotropic elastic media. Nevertheless, the local existence of the Rayleigh surface waves is enough to prove the same results as those for the isotropic case.

Key words. anisotropic elastic media, the Rayleigh surface waves, poles of the resolvent, local energy decay, the surface impedance tensor, trapping

AMS subject classifications. 35L20, 35P25, 35Q72, 73C35

PII. S0036141097314860

Introduction. It is well known that as a phenomenon of the propagation of the singularities, the Rayleigh surface wave which propagates along the boundary of an elastic body never penetrates into the body. One of our interests in this paper is to reconsider the Rayleigh surface wave as a phenomenon of the propagation of the energy. Especially, we want to prove that the energy of the Rayleigh surface wave penetrates into the body and the decay rate of the energy is slower than any negative power of time as time tends to infinity.

In order to state our results more precisely, let Ω be an exterior domain in $\mathbf{R}^n$ $(n \geq 3)$ with C^∞ and compact boundary Γ. We consider the domain Ω as a general anisotropic elastic medium with a traction free boundary. For the displacement vector $u(t,x) = {}^t(u_1(t,x), \dots, u_n(t,x))$, motions of the elastic medium are described by the following mixed problem:

$$\text{(0.1)} \qquad \begin{cases} (\partial_t^2 - A(x,\partial_x))u(t,x) = 0 & \text{in} \quad \mathbf{R} \times \Omega, \\ N(x,\partial_x)u(t,x) = 0 & \text{on} \quad \mathbf{R} \times \Gamma, \\ u(0,x) = f_1(x), \quad \partial_t u(0,x) = f_2(x) & \text{on} \quad \Omega. \end{cases}$$

In (0.1), the differential operator $A(x,\partial_x)$ is of the form

$$A(x,\partial_x)u = \sum_{i,l=1}^{n} \partial_{x_i}(C_{il}(x)\partial_{x_l}u),$$

*Received by the editors January 13, 1997; accepted for publication (in revised form) May 22, 1998; published electronically March 21, 2000.

http://www.siam.org/journals/sima/31-4/31486.html

†Faculty of Education, Ibaraki University, Mito, Ibaraki, 310-8512, Japan (kawasita @mito.ipc.ibaraki.ac.jp).

‡Department of Mathematics, Faculty of Engineering, Gunma University, Kiryu 376-8515, Japan (nakamura@math.sci.gunma-u.ac.jp).

and the boundary operator $N(x,\partial_x)$ is the conormal derivative of $A(x,\partial_x)$, which is represented as $N(x,\partial_x)u=\sum_{i,l=1}^{n}\nu_i(x)C_{il}(x)\partial_{x_l}u|_\Gamma$, where $\nu(x)={}^t(\nu_1(x),\nu_2(x),\dots,\nu_n(x))$ is the unit outer normal vector of Γ at $x\in\Gamma$.

The (j,k)-components $C_{ijkl}(x)$ of the $n\times n$-matrix $C_{il}(x)$ are called the elastic tensor. Due to the symmetry of $C_{ijkl}(x)$ given as assumption (A.1) below, the boundary condition in (0.1) describes the stress-free condition of the boundary.

It is well known that there is a surface wave called the Rayleigh surface wave (cf. Achenbach [1], Barnett and Lothe [2], Chadwick and Smith [3], Taylor [23]). Concerning the Rayleigh surface wave as a phenomenon of propagation of singularities, the singularities stay over the boundary and never go out from the boundary. Therefore, if the Rayleigh surface wave exists globally, it is "trapped" on the boundary.

In scattering theory, there is a principle asserting that the existence of trapped singularities of solutions yields poles of the resolvent and it also affects the behavior of the local energy decay of solutions. There are many works on this principle, especially for the acoustic wave equation (cf., e.g., Ikawa [5], Ralston [16], Vainberg [24]).

For isotropic elastic media, the Rayleigh surface waves exist globally in time. Hence, in this case, there are many works from the point of view in scattering theory (cf. [6], [10], [12], [18], [19], [20], [21], and [25]). Stefanov and Vodev [21] showed that there are poles of the resolvent approaching the real axis. However, for anisotropic elastic media, we do not know whether the Rayleigh surface waves exist globally in time even if they exist locally. Nevertheless, we can prove the same property for the resolvent. As a by-product of the proof, we can also prove that the local energy of solutions never decay at the rate of any negative power of time.

Throughout this paper, we suppose each component of the elastic tensor $C_{ijkl}(x)$ is of the form $C_{ijkl}(x)=C^0_{ijkl}+C^1_{ijkl}(x)$, where each C^0_{ijkl} is a real constant and $C^1_{ijkl}(x)$ is a real-valued C^∞ function on $\mathbf{R}^n$ with compact support. Denote by C^0_{il} (resp., $C^1_{il}(x)$) the matrices whose (j,k)-component is C^0_{ijkl} (resp., $C^1_{ijkl}(x)$). We set

$$A^0(\partial_x)u=\sum_{i,l=1}^{n}\partial_{x_i}(C^0_{il}\partial_{x_l}u),\qquad A^1(x,\partial_x)u=\sum_{i,l=1}^{n}\partial_{x_i}(C^1_{il}(x)\partial_{x_l}u).$$

Note that $A(x,\partial_x)u=A^0(\partial_x)u+A^1(x,\partial_x)u$.

Further, we always assume the following physically natural assumptions:

$$\text{(A.1)}\qquad C_{ijkl}(x)=C_{jikl}(x)=C_{lkji}(x)\quad\text{for any } x\in\mathbf{R}^n \text{ and } i,j,k,l=1,\dots,n;$$

(A.2) there is a constant $\delta>0$ such that

$$\sum_{i,j,k,l=1}^{n}C_{ijkl}(x)\epsilon_{kl}\bar{\epsilon}_{ij}\ge\delta\sum_{i,j=1}^{n}|\epsilon_{ij}|^2$$

for any $x\in\mathbf{R}^n$ and $n\times n$-symmetrix matrix (ϵ_{ij}).

From (A.1) and (A.2), we can define the outgoing (resp., incoming) resolvent $R^+(z)$ (resp., $R^-(z)$) as follows. The resolvent $R^\pm(z)$ of the stationary problem

$$\text{(0.2)}\qquad\begin{cases}(A(x,\partial_x)+z^2)v(x;z)=f(x) & \text{in}\quad\Omega,\\ N(x,\partial_x)v(x;z)=0 & \text{on}\quad\Gamma\end{cases}$$

is a $B(L^2(\Omega),\ H^2(\Omega))$-valued holomorphic function in $\pm\text{Im}\ z < 0$. For any $a > 0$ with $\Gamma \subset B_a = \{x \in \mathbf{R}^n\,;\ |x| < a\}$, $R^{\pm}(z)$ can be continued meromorphically as a $B(L^2_a(\Omega), H^2(\Omega_a))$-valued function on $\widetilde{\mathbf{C}}_{\pm}$, where $L^2_a(\Omega) = \{f \in L^2(\Omega)\,;\ f(x) = 0$ in $|x| > a\}$, $\Omega_a = \Omega \cap B_a$ and $\widetilde{\mathbf{C}}_{\pm} = \mathbf{C}$ if n is odd and $\widetilde{\mathbf{C}}_{+} = \{z \in \mathbf{C} \setminus \{0\}\,;\ -3\pi/2 < \arg z < \pi/2\}$, $\widetilde{\mathbf{C}}_{-} = \{|z|\exp(-\sqrt{-1}\arg z)\,;\ z \in \widetilde{\mathbf{C}}_{+}\}$ if n is even (see Appendix A). This extended operator $R^+(z)$ (resp., $R^-(z)$) is called the outgoing (resp., incoming) resolvent and $v^+(x; z) = R^+(z)f(x)$ (resp., $v^-(x; z) = R^-(z)f(x)$) is the outgoing (resp., incoming) solution of (0.2).

Unlike the isotropic case, the Rayleigh surface waves do not always exist. The necessary and sufficient condition (ERW) for their existence, i.e., existence of the Rayleigh surface waves, is given in terms of the surface impedance tensor. If the space dimension is three, the condition (ERW) coincides with the condition given in Nakamura [15]. The condition (ERW) was first introduced by Barnett and Lothe [2] for the homogeneous (i.e., $C^1_{il}(x) = 0$) anisotropic elastic medium with flat boundaries. The precise definition of the condition (ERW) and the surface impedance tensor are given in section 1.

The first purpose of this paper is to show that the existence of the Rayleigh surface waves affects the location of the poles of the resolvent. Concerning this, we have the following theorem.

Theorem 0.1. *Assume that* (A.1) *and* (A.2) *hold and that the space dimension* n *is odd. If the Rayleigh surface waves appear, that is, the condition* (ERW) *in section* 1 *is satisfied, then for any constants* C_0, $C_1 > 0$ *and integer* $N > 0$, *the resolvent* $R^{\pm}(z)$ *has poles in the region* $0 \leq \pm\text{Im}\ z \leq C_0|\text{Re}\ z|^{-N}$, $|\text{Re}\ z| \geq C_1$.

Note that the condition (ERW) is automatically satisfied for the isotropic case. Furthermore, we know the explicit form of the principal symbol of the Neumann operator in the elliptic region. In [21], they used this fact essentially in proving the result corresponding to Theorem 0.1 for the isotropic case. However, the condition (ERW) only gives the fact that the principal symbol of the Neumann operator is locally simple characteristic in the elliptic region. Nevertheless, this is enough to show Theorem 0.1.

From Theorem 0.1, we can show in the next corollary the existence of the poles of the resolvents approaching the real axis.

Corollary 0.2. *Under the same assumption as in Theorem* 0.1, *there exists a sequence of distinct poles* z_j $(j = 1, 2, \dots)$ *of* $R^{\pm}(z)$ *satisfying*

$$0 \leq \pm\text{Im}\ z_j \leq C_N|\text{Re}\ z_j|^{-N}, \qquad (j = 1, 2, \dots)\ \textit{for all}\ N \in \mathbf{N}.$$

In the homogeneous case (i.e., $C^1_{il}(x) = 0$), as in Iwashita and Shibata [9], Iwashita [8], if we further assume that the characteristic roots of $A^0(\xi) = \sum_{i,l=1}^n C^0_{il}\xi_i\xi_l$ are of constant multiplicity for all $\xi \in \mathbf{R}^n \setminus \{0\}$, the resolvent is holomorphic on the real axis except the origin. Note that in the homogeneous isotropic case (i.e., $C^1_{il}(x) = 0$ and $C^0_{ijkl} = \lambda^0\delta_{ij}\delta_{lk} + \mu^0(\delta_{il}\delta_{jk} + \delta_{ik}\delta_{jl})$) these additional assumptions are also satisfied.

The second purpose of this paper is to consider the asymptotic behavior of the local energy. We define the local energy of the solution $u(t, x)$ of (0.1) at time t in a region D as

$$\mathbf{E}(u, D, t) = \frac{1}{2}\int_D \left\{ \sum_{i,j,k,l=1}^n C_{ijkl}(x)\partial_{x_l}u_k(t, x)\overline{\partial_{x_i}u_j(t, x)} + |\partial_t u(t, x)|^2 \right\} dx.$$

For any $a > 0$ with $\Gamma \subset B_a$ and integer $m \geq 0$, we set

$$p_{m,a}(t) = \sup \left\{ \frac{\mathbf{E}(u, \Omega_a, t)}{\left\|\nabla_x f_1\right\|^2_{H^m(\Omega)} + \left\|f_2\right\|^2_{H^m(\Omega)}} \, ; \, 0 \neq f \in C_0^\infty(\overline{\Omega} \cap B_a) \right\},$$

where $u(t, x)$ is the solution of problem (0.1) with the initial data $f = {}^t(f_1, f_2)$. We call $p_{m,a}(t)$ the uniform decay rate of the local energy. Our second purpose is to show that the existence of the Rayleigh surface waves also affects the behavior of the uniform decay rate.

For acoustic waves, Ralston [16] shows that if there exists a trapping ray of geometrical optics (i.e., the existence of trapping singularities of solutions of the acoustic wave equation), the uniform decay rate $p_{0,a}(t)$ in the sense of Morawetz [14] never tends to 0 as $t \to \infty$. For isotropic elastic media, as in [10] and [11], we have the same conclusion as in the acoustic case because there exist trapping singularities due to the global existence of the Rayleigh surface waves.

For $p_{m,a}(t)$ with $m \geq 1$, however, by the argument of Walker [26], we can show $\lim_{t\to\infty} p_{m,a}(t) = 0$ even in the general homogeneous anisotropic case. In fact, from (A.1) and (A.2), the proof given by Shibata and Soga [17] implies the local energy decay property; that is, $\lim_{t\to\infty} \mathbf{E}(u, D, t) = 0$ if D is bounded. This property and Rellich compactness theorem give the basis on Walker's proof.

In the isotropic case, Ikehata and Nakamura [6] showed that for any $\alpha > 0$ and $m \geq 0$, we cannot have an estimate of the form $p_{m,a}(t) \leq C \exp(-\alpha t)$ if Γ is the unit sphere in $\mathbf{R}^3$. For general boundary, as in [12], the uniform decay rate $p_{m,a}(t)$ ($m \geq 1$) never allows an estimate from above by any negative power of time.

Even in the general anisotropic case, we obtain the same conclusions as in the isotropic case.

THEOREM 0.3. *We assume that* (A.1) *and* (A.2) *are satisfied. If we further assume the condition* (ERW) *holds, then for any $a > 0$ with $\Gamma \subset B_a$, we have*

(1) $\lim_{t\to\infty} p_{0,a}(t) \neq 0$,

(2) $\lim_{t\to\infty} p_{m,a}(t) t^\gamma = \infty$ *for any $\gamma > 0$,* $m \in \mathbf{N} \cup \{0\}$.

Thus, although we only assume the local existence condition (ERW) of the Rayleigh surface waves, we have such conclusions for $p_{m,a}(t)$. This is quite surprising, because the local existence of the Rayleigh surface waves does not mean that there exist trapping singularities. It affects the local energy decay property of the solution which is a global property of the solution.

1. Existence of the Rayleigh surface waves and the surface impedance tensor.

Here we review the definition of the surface impedance tensor and its relationship to the existence of the Rayleigh surface waves. For the hypersurface $\Gamma \subset \mathbf{R}^n$, the restriction to Γ of the cotangent bundle $T^*(\mathbf{R}^n)$ has the orthogonal decomposition $T^*(\mathbf{R}^n)|_\Gamma = N^*(\Gamma) \oplus (N^*(\Gamma))^\perp$ by the Euclidian metric on $\mathbf{R}^n$, where $N^*(\Gamma)$ is the conormal bundle of $\Gamma \subset \mathbf{R}^n$. Note that the canonical map $(N^*(\Gamma))^\perp \ni \zeta \mapsto \zeta|_{T^*(\Gamma)} \in T^*(\Gamma)$ is isometric, where the fiber metric $\left\|\zeta\right\|_\Gamma$ of the cotangent bundle $T^*(\Gamma)$ is the one induced by the Euclidian metric on $\mathbf{R}^n$. We regard the unit outer normal vector $\nu(x)$ as a C^∞-section of the conormal bundle $N^*(\Gamma)$ by $\sum_{j=1}^n \nu_j(x) dx_j$.

(i) Elliptic region and limiting velocity. Let $s : \widetilde{U} \ni \sigma = {}^t(\sigma_1, \ldots, \sigma_{n-1}) \mapsto s(\sigma) \in U \subset \Gamma$ be a local coordinates system, where $\widetilde{U} \subset \mathbf{R}^{n-1}$ is an open set. The coordinates of $\mathbf{R} \times U$ are given by $\kappa : \mathbf{R} \times \widetilde{U} \ni (t, \sigma) \mapsto (t, s(\sigma)) \in \mathbf{R} \times U$. $\tilde{s} : T^*(\widetilde{U}) \to T^*(U)$

and $\tilde{\kappa} : T^*(\mathbf{R} \times \widetilde{U}) \to T^*(\mathbf{R} \times U)$ denote the local triviality of $T^*(\Gamma)$ and $T^*(\mathbf{R} \times \Gamma)$, respectively.

Definition 1.1. *The elliptic region $\mathcal{E} \subset T^*(\mathbf{R} \times \Gamma) \setminus \{0\}$ of the operator $\partial_t^2 - A(x, \partial_x)$ and the elliptic region $\mathcal{E}_\Gamma \subset T^*(\Gamma)$ of the operator $A(x, \partial_x) + z^2$ are defined by*

$$\begin{aligned}(\tilde{\kappa})^{-1}\mathcal{E} = \{(t, \sigma, \tau, \eta) \in T^*(\mathbf{R} \times \widetilde{U}) \ ; \ &\det(\tau^2 I - \sigma_p(A)(s(\sigma), \theta(\sigma, \eta, q))) = 0 \\ &\textit{has no real root as a polynomial in } q\,\}\end{aligned}$$

and

$$(\tilde{s})^{-1}\mathcal{E}_\Gamma = \{(\sigma, \eta) \in T^*(\widetilde{U}) \ ; \ (t, \sigma, 1, \eta) \in \tilde{\kappa}^{-1}(\mathcal{E}) \quad \textit{for some } t \in \mathbf{R}\},$$

respectively. Here $\sigma_p(A)(x, \xi) = \sum_{i,l=1}^n C_{il}(x)\xi_i\xi_l$ is the principal symbol of the operator $A(x, \partial_x)$ and the vector $\theta = {}^t(\theta_1, \dots, \theta_n) = \theta(\sigma, \eta, q)$ is defined by $\sum_{j=1}^n \theta_j dx_j = \sum_{j=1}^{n-1} \eta_j d\sigma_j + q \sum_{j=1}^n \nu_j(s(\sigma))dx_j$ in $(T^*(\mathbf{R}^n)|_\Gamma)^{\mathbf{C}}$, i.e., the complexification of the vector bundle $T^*(\mathbf{R}^n)|_\Gamma$.

Note that the elliptic region $\mathcal{E}$ (resp., $\mathcal{E}_\Gamma$) is well defined as an open conic subset in $T^*(\mathbf{R} \times \Gamma)$ (resp., an open subset in $T^*(\Gamma)$).

Definition 1.2. *The limiting velocity $v_L(\zeta)$ at $\zeta \in T^*(\Gamma) \setminus \{0\}$ is defined by*

$$(\tilde{s}^* v_L)(\sigma, \eta) = \inf\{c_i(\sigma, \eta, \phi)(\cos\phi)^{-1} \ ; \ i = 1, 2, \dots, n, |\phi| < \pi/2\},$$

where $c_i(\sigma, \eta, \phi)$ $(i = 1, 2, \dots, n)$ are the positive real roots of the polynomial in c

$$\det(c^2 I - \sigma_p(A)(s(\sigma), \tilde{\theta}(\sigma, \eta, \phi))) = 0$$

with $\tilde{\theta}(\sigma, \eta, \phi) = (\cos\phi)\theta(\sigma, \eta, 0)|\eta|_\Gamma^{-1} + (\sin\phi)\nu(s(\sigma))$, $|\eta|_\Gamma = \left\|s((\sigma, \eta))\right\|_\Gamma$.

Note that the limiting velocity v_L is well defined as a continuous function on $T^*(\Gamma) \setminus \{0\}$, homogeneous of order 0, $(\tilde{s}^* v_L)(\sigma, -\eta) = (\tilde{s}^* v_L)(\sigma, \eta)$ for any $(\sigma, \eta) \in T^*(\widetilde{U})$. By Definitions 1.1 and 1.2, we have $\mathcal{E} = \{(t, \tau, \zeta) \in T^*(\mathbf{R}) \times T^*(\Gamma) \ ; \ |\tau| < v_L(\zeta) \left\|\zeta\right\|_\Gamma\}$ and $\mathcal{E}_\Gamma = \{\zeta \in T^*(\Gamma) \ ; \ 1 < v_L(\zeta) \left\|\zeta\right\|_\Gamma\}$.

(ii) Surface impedance tensor. Now we introduce the surface impedance tensor. We fix $(\sigma, \eta) \in T^*(\Gamma)$, $\tau \in \mathbf{R}$ with $(\sigma, \eta/\tau) \in \mathcal{E}_\Gamma$ and set $\tilde{\nu} = -\nu(s(\sigma))$, $m = \theta(\sigma, \eta, 0)|\eta|_\Gamma^{-1}$, $v = \tau/|\eta|_\Gamma$. Noting that $|m|^2 = \sum_{j=1}^n m_j^2 = 1$, $m \cdot \tilde{\nu} = 0$, we recall the Stroh formalism in [2], [3], [15], and [22] used to define the surface impedance tensor. We consider solutions $u(t, x)$ of equation $(\partial_t^2 - A(s(\sigma), \partial_x))u(t, x) = 0$ of the form $u(t, x) = a\exp(\sqrt{-1}k(m \cdot x + p\tilde{\nu} \cdot x - vt))$ with positive parameter k. We seek such solutions decaying exponentially in $\tilde{\nu} \cdot x$, which is equivalent to finding p with $\operatorname{Im} p > 0$ and $a \neq 0$ satisfying $(v^2 - A(s(\sigma), m + p\tilde{\nu}))a = 0$. Since $|v| < \tilde{s}^* v_L(\sigma, \eta)$, we always have such a pair of p and a. For them, we set $l = -\sum_{i,l=1}^n \tilde{\nu}_i C_{il}(s(\sigma))(m_l + p\tilde{\nu}_l)a$.

Setting $\xi = {}^t({}^t a, {}^t l) \in \mathbf{C}^{2n}$, we can seek p and these vectors a and l by the so called Stroh's eigenvalue problem $(N - p)\xi = 0$, where

$$\begin{aligned}&N = \\ &-\begin{pmatrix} \langle\tilde{\nu}, \tilde{\nu}\rangle^{-1}\langle\tilde{\nu}, m\rangle & \langle\tilde{\nu}, \tilde{\nu}\rangle^{-1} \\ \langle m, \tilde{\nu}\rangle\langle\tilde{\nu}, \tilde{\nu}\rangle^{-1}\langle\tilde{\nu}, m\rangle - \langle\tilde{\nu}, \tilde{\nu}\rangle^{-1} & \langle m, \tilde{\nu}\rangle^{-1}\langle\tilde{\nu}, \tilde{\nu}\rangle^{-1} \end{pmatrix}\end{aligned}$$

with $\langle b, \tilde{b}\rangle = \sum_{i,l=1}^n (C_{il}(s(\sigma)) - v^2 m_i m_l)b_i\tilde{b}_l = \sum_{i,l=1}^n (C_{il}(s(\sigma)) - \tau^2|\eta|_\Gamma^{-4}\, \theta_i(\sigma, \eta, 0)\theta_l(\sigma, \eta, 0))b_i\tilde{b}_l$ for $b = {}^t(b_1, \dots, b_n)$, $\tilde{b} = {}^t(\tilde{b}_1, \dots, \tilde{b}_n)$.

Since $|v| < \tilde{s}^* v_L(\sigma, \eta)$, N has $2n$ nonreal eigenvalues p_α ($\alpha = 1, 2, \ldots, 2n$) accounting with multiplicity. We order them so that $\mathrm{Im}\, p_\alpha > 0$ and $p_{\alpha+n} = \overline{p_\alpha}$ (for $\alpha = 1, 2, \ldots, n$). We take generalized eigenvectors ξ_α ($\alpha = 1, 2, \ldots, 2n$) corresponding to the eigenvalues p_α. We write the first n-components of ξ_α as a_α and the last n-components of ξ_α as l_α, so that $\xi_\alpha = {}^t({}^t a_\alpha, {}^t l_\alpha)$. Then, as in [2], [3], and [15], there is an $n \times n$-matrix $Z = Z(\sigma, \eta, \tau)$ such that $l_\alpha = \sqrt{-1} Z a_\alpha$ ($\alpha = 1, 2, \ldots, n$) hold. This matrix is called the surface impedance tensor.

The surface impedance tensor is well defined as an $n \times n$-matrix valued function in $\mathcal{E}$ independent of time t and is homogeneous of order 0 in (σ, η). In fact, as in [2], [3], and [15], we can represent $Z(\sigma, \eta, \tau)$ as follows:

$$Z(\sigma,\eta,\tau) = -(Q^{-1}(\sigma,\eta,\tau) + \sqrt{-1} Q^{-1}(\sigma,\eta,\tau) S(\sigma,\eta,\tau)),$$

where

$$S(\sigma,\eta,\tau) = -(2\pi)^{-1} \int_0^{2\pi} \langle w(\phi), w(\phi)\rangle^{-1} \langle w(\phi), \eta(\phi)\rangle \, d\phi,$$
$$Q(\sigma,\eta,\tau) = -(2\pi)^{-1} \int_0^{2\pi} \langle w(\phi), w(\phi)\rangle^{-1} \, d\phi,$$

$w(\phi) = -(\sin\phi)\theta(\sigma, \eta/|\eta|_\Gamma, 0) + (\cos\phi)(-\nu(s(\sigma)))$, $\eta(\phi) = (\cos\phi)\theta(\sigma, \eta/|\eta|_\Gamma, 0) + (\sin\phi)$ $(-\nu(s(\sigma)))$.

We also denote by $Z(\zeta, \tau)$ ($\zeta \in T^*(\Gamma)$, $\tau \in \mathbf{R}$ with $(t, \tau, \zeta) \in \mathcal{E}$ for some $t \in \mathbf{R}$) the surface impedance tensor defined globally in $\mathcal{E}$. By the facts in [2] and [3], the surface impedance tensor $Z(\zeta, \tau)$ is a Hermite matrix for any (ζ, τ) and has a limit as τ tends to $\|\zeta\|_\Gamma v_L(\zeta)$, which is also denoted by $Z(\zeta, \|\zeta\|_\Gamma v_L(\zeta))$ hereafter; that is, Z is continuous in $\overline{\mathcal{E}} \subset T^*(\mathbf{R} \times \Gamma)\{0\}$.

(iii) Condition (ERW). Now we state the condition (ERW) ensuring the existence of the Rayleigh surface waves.

(ERW) There exists a point $\zeta^0 \in T^*(\Gamma)$ such that the Hermite matrix $Z(\zeta^0, v_L(\zeta^0) \|\zeta^0\|_\Gamma)$ is not nonnegative definite.

In the three dimensional case, it is the same conditon as in Theorem 12 of Barnett and Lothe [2]. Nakamura [15] shows the existence of the Rayleigh surface waves in the sense of propagation of singularities by deducing a good property for the principal symbol of the (time dependent) Neumann operator from the condition in Barnett and Lothe [2] (cf. Theorems 2.2, 2.3, and 2.4 in [15]). By the historical reasons above, we adopt the condition (ERW) as a condition describing the existence of the Rayleigh surface waves.

2. Outline of proof. In this section, we give our plan to show Theorems 0.1 and 0.3. Both theorems are shown by contradiction arguments. In both cases, the denial of theorems implies estimates of the resolvent near the real axis. Eventually, this fact becomes inconsistent to the assumption (ERW).

For Theorem 0.1, by an argument similar to the proof of Proposition 1 in Stefanov and Vodev [21], we can obtain the following proposition.

Proposition 2.1. *Assume that* (A.1) *and* (A.2) *hold and that n is odd. If Theorem* 0.1 *is not true, that is, the resolvent $R^{\pm}(z)$ is holomorphic in the region* $|\mathrm{Im}\, z| \le C|\mathrm{Re}\, z|^{-m_0}$, $|\mathrm{Re}\, z| \ge C'$ *with some fixed constants $C, C' > 0$ and $m_0 > 0$,* *then there exist constants* $C_0, C_1, C_2 > 0$ *such that*

$$\begin{aligned} &\left\|R^{\pm}(z)f\right\|_{H^2(\Omega_a)} \le C_2|z|^{m_0+3n+5}\left\|f\right\|_{L^2(\Omega)} \\ &\text{for any } f \in L^2_a(\Omega), z \in \widetilde{\mathbf{C}}_{\pm}, |\mathrm{Im}\ z| \le C_0|\mathrm{Re}\ z|^{-(m_0+3n+4)}, |\mathrm{Re}\ z| \ge C_1. \end{aligned} \tag{2.1}$$

For Theorem 0.3, if it is not true, we have an estimate $p_{m,a}(t) \le Ct^{-\gamma}$ with some fixed constants $C, \gamma > 0$, $m \in \mathbf{N}$ and $a > 0$ satisfying $\Gamma \subset B_a$. Then, like the proof of Lemma 7.3 in [12], we can prove an estimate like (2.1). Therefore, in both cases, it only suffices to derive a contradiction by assuming the estimate (2.1) and the condition (ERW).

We argue as follows. First, in the elliptic region, we approximate the Neumann operator by a pseudodifferential operator $B_N(z)$ with (complex) parameter z. The operator $B_N(z)$ is nonelliptic due to the condition (ERW) (cf. sections 4 and 5).

Second, we construct a pseudodifferential operator $P(z)$ with a symbol defined on the whole $T^*(\Gamma)$ such that $P(z)$ is an extension of $z^{2N-1}B_{2N}(z)$ and it is elliptic in $T^*(\Gamma)$ except where $B_{2N}(z)$ is nonelliptic.

Third, we show the existence of a sequence z_j $(j = 1, 2, \dots)$ with the properties that $P(z_j)$ has null solutions and $\lim_{j\to\infty} \mathrm{Re}\ z_j = \infty$. This is done by showing a contradiction under the assumption such that these z_j $(j = 1, 2, \dots)$ do not exist. We start with preparing a priori estimates of the Neumann operator $T^{\pm}(z)$ and the resolvent (cf. section 3). Using these a priori estimates and the fact that outgoing and incoming Neumann operators are essentially the same in the elliptic region, we have a priori estimates of $B_{2N}(z)$ in $\mathrm{Im}\ z \ne 0$. These estimates imply the invertibility of $P(z)$ for z with sufficiently large $|\mathrm{Im}\ z|$ and $\mathrm{Re}\ z$. Then, from the denial of the existence of z_j, the same argument as in [21] implies the estimates $\left\|f\right\|_{L^2(\Gamma)} \le C|z|^{-2N+1}(\log z)^{-1}\left\|P(z)f\right\|_{L^2(\Gamma)}$ for any $f \in C^\infty(\Gamma)$ and real $z >> 1$ (cf. (7.11) in section 7). Then the condition (ERW) allows us to construct an approximation solution of $P(z)$ which breaks the above estimate (cf. section 7).

Last, we take null solutions f_j of $P(z_j)$ with $\left\|f_j\right\|_{L^2(\Gamma)} = 1$ $(j = 1, 2, \dots, n)$. Then the functions f_j are asymptotically null solutions of $B_{2N}(z_j)$ (cf. Theorem 6.1). On the other hand, by the estimate (2.1), we can show an estimate of $B_{2N}(z)$ (cf. (6.1) in section 6). However, this estimate is inconsistent with the existence of the asymptotic null solutions of $B_{2N}(z_j)$ if N is sufficiently large (cf. section 6).

The prototype of this procedure is proposed by Stefanov and Vodev [21] to show the existence of the poles of the resolvent in the isotropic case. In this case, we know the form of the principal symbol of the Neumann operator. The arguments in [21] are based on this fact. In the anisotropic case, however, we cannot know such a global structure of the principal symbol of the Neumann operator. Hence, this is the main difficulty in the anisotropic case which requires a new idea.

From the condition (ERW), we can only show that the principal symbol of the Neumann operator is "real principal type" locally (cf. Proposition 5.2). This is the reason why we can only know the local existence of the Rayleigh surface waves from the condition (ERW). In our argument, however, this is enough to show results analogous to those in the isotropic case. Furthermore, our procedure is quite general, which can be applied to other equations.

3. The Neumann operator. We start by introducing the Neumann operator $T^{\pm}(z)$ for the reduced problem which plays a crucial role in proving the main theorems.

We consider the reduced elastic wave equation with an inhomogeneous Dirichlet datum $g(x)$,

$$(3.1)\qquad \begin{cases} (A(x,\partial_x)+z^2)v^\pm(x;z)=0 & \text{in} \quad \Omega, \\ v^\pm(x;z)=g(x) & \text{on} \quad \Gamma, \\ \ v^+(x;z) \text{ is outgoing (resp., } v^-(x;z) \text{ is incoming),} & \end{cases}$$

where "outgoing" and "incoming" are defined as similar manners as in the Introduction. We denote by $U^+(z)$ (resp., $U^-(z)$) the outgoing (resp., incoming) solution operator of (3.1). The Neumann operator $T^\pm(z)$ is defined as

$$T^\pm(z)g(x) = N(x,\partial_x)U^\pm(z)g(x)|_\Gamma.$$

Since the operator $U^+(z)$ (resp., $U^-(z)$) can be represented by the outgoing (resp., incoming) resolvent of the reduced problem for mixed problem with Dirichlet boundary condition, the Neumann operator $T^\pm(z)$ is a $B(H^{3/2}(\Gamma), H^{1/2}(\Gamma))$-valued holomorphic function in $\pm\mathrm{Im}\ z<0$, and it can be continued meromorphically in $\widetilde{\mathbf{C}}_\pm$. Moreover, $(T^\pm(z))^{-1}$ is a $B(H^{1/2}(\Gamma), H^{3/2}(\Gamma))$-valued holomorphic function in $\pm\mathrm{Im}\ z<0$ and meromorphic function in $\widetilde{\mathbf{C}}_\pm$, since $(T^\pm(z))^{-1}$ is represented by $R^\pm(z)$.

Lemma 3.1. *Assume that* (A.1) *and* (A.2) *hold. If the resolvent satisfies the estimate* (2.1)*, the inverse* $(T^\pm(z))^{-1}$ *of the Neumann operator is holomorphic in* $|\mathrm{Im}\ z| \le C_0|\mathrm{Re}\ z|^{-(m_0+3n+4)}$, $|\mathrm{Re}\ z|\ge C_1$ *and there is a constant* $C_3>0$ *such that*

$$\begin{aligned} &\|(T^\pm(z))^{-1}\|_{L^2(\Gamma)} \le C_3|z|^{m_0+3n+7}\,\|f\|_{L^2(\Gamma)} \\ &\qquad \textit{for any } f \in C^\infty(\Gamma), |\mathrm{Im}\ z|\le C_0|\mathrm{Re}\ z|^{-(m_0+3n+4)}, |\mathrm{Re}\ z|\ge C_1, \end{aligned}$$

where $C_0, C_1>0$ *are the same constant as in Proposition* 2.1.

Proof. From the relation between $(T^\pm(z))^{-1}$ and $R^\pm(z)$, we have the holomorphicity of $(T^\pm(z))^{-1}$ and the estimate

$$\begin{aligned} &\|(T^\pm(z))^{-1}\|_{B(H^{1/2}(\Gamma))} \le C|z|^{m_0+3n+7} \\ &\qquad \text{in } |\mathrm{Im}\ z| \le C_0|\mathrm{Re}\ z|^{-(m_0+3n+4)}, |\mathrm{Re}\ z|\ge C_1. \end{aligned}$$

Noting the duality relation $((T^+(z))^{-1}g,h)_{L^2(\Gamma)} = (g,(T^-(\overline{z}))^{-1}h)_{L^2(\Gamma)}$ and the following property of the Sobolev norm

$$\|g\|_{H^s(\Gamma)} \le C(s)\sup\{|(g,h)_{L^2(\Gamma)}|(\|h\|_{H^{-s}(\Gamma)})^{-1}\,;\, 0\ne h\in C^\infty(\Gamma)\},$$

we obtain the estimate

$$\begin{aligned} &\|(T^\pm(z))^{-1}\|_{B(H^{-1/2}(\Gamma))} \le C|z|^{m_0+3n+7} \\ &\qquad \text{in } |\mathrm{Im}\ z| \le C_0|\mathrm{Re}\ z|^{-(m_0+3n+4)}, |\mathrm{Re}\ z|\ge C_1; \end{aligned}$$

(cf. section 2 in [12]). Hence, well-known interpolation results give us Lemma 3.1.

We also need the following lemma.

Lemma 3.2. *If we assume* (A.1) *and* (A.2) *are satisfied, there exists a constant* $C>0$ *such that for any* $\delta>0$ *we have*

$$\begin{aligned} &\|(T^\pm(z))^{-1}\|_{B(L^2(\Gamma))} \le C(1+\delta)|\mathrm{Im}\ z|^{-1}, \\ &\|R^\pm(z)\|_{B(L^2(\Omega),H^1(\Omega))} \le C(1+\delta)^{1/2}|\mathrm{Im}\ z|^{-1} \\ &\textit{for any } z\in\widetilde{\mathbf{C}}_\pm, \pm\mathrm{Im}\ z<0, |\mathrm{Im}\ z|\le\delta|\mathrm{Re}\ z|, \mathrm{Re}\ z\ge 1. \end{aligned}$$

Proof. For a function $g \in C^\infty(\Gamma)$ and $z \in \widetilde{\mathbf{C}}_\pm$, $\pm\mathrm{Im}\ z < 0$, we take the outgoing (resp., incoming) solution $w^+(x;z)$ (resp., $w^-(x;z)$) of the problem

$$\begin{cases} (A(x,\partial_x) + z^2)w^\pm(x;z) = 0 & \text{in} \quad \Omega, \\ N(x,\partial_x)w^\pm(x;z) = g(x) & \text{on} \quad \Gamma, \end{cases}$$

respectively. Since $w^\pm(x;z)$ is a L^2-solution, integration by parts gives

$$\begin{aligned} \sum_{i,l=1}^{n} (C_{il}(\cdot)\partial_{x_l} w^\pm(\cdot;z), \partial_{x_i} w^\pm(\cdot;z))_{L^2(\Omega)} + (\mathrm{Im}\ z)^2 \left\| w^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} \\ = (\mathrm{Re}\ z)^2 \left\| w^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} + \mathrm{Re}\ (g, w^\pm(\cdot;z))_{L^2(\Gamma)}, \end{aligned} \tag{3.2}$$

$$2\mathrm{Re}\ z \mathrm{Im}\ z \left\| w^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} = -\mathrm{Im}\ (g, w^\pm(\cdot;z))_{L^2(\Gamma)}. \tag{3.3}$$

Recall the well-known estimate, that is, for any $a > 0$ with $\Gamma \subset B_a$, there exists $C_a > 0$ such that

$$\begin{aligned} \left\| v|_\Gamma \right\|^2_{L^2(\Gamma)} \leq & C_a\{\epsilon \left\| \nabla_x v \right\|^2_{L^2(\Omega_a)} + \epsilon^{-1} \left\| v \right\|^2_{L^2(\Omega_a)}\} \\ & \text{for any } \ 0 < \epsilon \leq 1, v \in H^1(\Omega_a). \end{aligned}$$

For $z \in \widetilde{\mathbf{C}}_\pm$ with $\pm\mathrm{Im}\ z < 0$, we use the estimate for $\epsilon = |\mathrm{Re}\ z|^{-1}$, $v = w^\pm(\cdot;z)$, which yields

$$\begin{aligned} |\mathrm{Re}\ z| \left\| w^\pm(\cdot;z)|_\Gamma \right\|^2_{L^2(\Gamma)} \leq C_a \Big\{ & \left\| \nabla_x w^\pm(\cdot;z) \right\|^2_{L^2(\Omega_a)} \\ & + |\mathrm{Re}\ z|^2 \left\| w^\pm(\cdot;z) \right\|^2_{L^2(\Omega_a)} \Big\}. \end{aligned} \tag{3.4}$$

The estimate (3.4), the equalities (3.2), (3.3), and the Korn's inequality (cf. Shibata and Soga [17] or Ito [7]) imply

$$|\mathrm{Re}\ z| \left\| w^\pm(\cdot;z)|_\Gamma \right\|^2_{L^2(\Gamma)} \leq C_a C \left(1 + \frac{|\mathrm{Re}\ z|}{|\mathrm{Im}\ z|} \right) |(g, w^\pm(\cdot;z))_{L^2(\Gamma)}|$$

with some fixed constant $C > 0$. Since $(T^\pm(z))^{-1}g(x) = w^\pm(\cdot;z)|_\Gamma$, the estimate gives us Lemma 3.2 for $(T^\pm(z))^{-1}$.

Next we turn to show the estimate of $R^\pm(z)$. Putting $v^\pm(x;z) = R^\pm(z)f(x)$ and integrating by parts we obtain

$$\begin{aligned} \sum_{i,l=1}^{n} (C_{il}(\cdot)\partial_{x_l} v^\pm(\cdot;z), \partial_{x_i} v^\pm(\cdot;z))_{L^2(\Omega)} + (\mathrm{Im}\ z)^2 \left\| v^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} \\ = (\mathrm{Re}\ z)^2 \left\| v^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} - \mathrm{Re}\ (f, v^\pm(\cdot;z))_{L^2(\Omega)}, \\ 2\mathrm{Re}\ z \mathrm{Im}\ z \left\| v^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} = \mathrm{Im}\ (f, v^\pm(\cdot;z))_{L^2(\Omega)} \end{aligned}$$

for $z \in \widetilde{\mathbf{C}}_\pm$, $\pm\mathrm{Im}\ z < 0$. From the equalities above, it follows that

$$\begin{aligned} \left\| \nabla_x v^\pm(\cdot;z) \right\|^2_{L^2(\Omega)} \leq C & \left\| v^\pm(\cdot;z) \right\|_{L^2(\Omega)} \\ & \Big\{ (\mathrm{Re}\ z)^2 \left\| v^\pm(\cdot;z) \right\|_{L^2(\Omega)} + \left\| f \right\|_{L^2(\Omega)} \Big\}, \\ 2|\mathrm{Re}\ z \mathrm{Im}\ z| \left\| v^\pm(\cdot;z) \right\|_{L^2(\Omega)} \leq & \left\| f \right\|_{L^2(\Omega)}, \end{aligned}$$

where we use the Korn's inequality to deduce the estimate of $\nabla_x v^{\pm}(\cdot; z)$. The estimate implies Lemma 3.2 for $R^{\pm}(z)$.

4. Approximation operators of the Neumann operator. In this section, we construct an approximate operator of the Neumann operator in the elliptic region $\mathcal{E}_\Gamma$ via constructing an approximation of the Poisson operator $U^{\pm}(z)$. Let $s^k : \widetilde{U}_k \ni \sigma = (\sigma_1, \dots, \sigma_{n-1}) \mapsto s^k(\sigma) \in U_k \subset \Gamma$ $(k = 1, 2, \dots, N_0)$ be a local coordinates system satisfying $\Gamma = \cup_{k=1}^{N_0} U_k$. For s^k, we denote by $\widetilde{s}^k$ the local triviality of $T^*(U_k)$. For sufficiently small $b_0 > 0$, the functions $x^k(r, \sigma) = s^k(\sigma) - r\nu(s^k(\sigma))$ defined in $|r| \le b_0$, $\sigma \in \widetilde{U}_k (k = 1, \dots, N_0)$ give us a local coordinate system in $\mathbf{R}^n$ near Γ.

To begin, we construct a cutoff operator $X(z)$ in the elliptic region. Take functions $\phi_k(x) \in C_0^\infty(U_k)$ $(k = 1, \dots, N_0)$ such that $\sum_{k=1}^{N_0} (\phi_k(x))^2 = 1$ on Γ. For fixed open sets $W_0 \subset W_1 \subset \mathcal{E}_\Gamma$ with $\overline{W_0} \subset W_1$, $\overline{W_1} \subset \mathcal{E}_\Gamma$, which are chosen in Lemma 5.1 in section 5 precisely, we take functions $a(\zeta) \in C^\infty(T^*(\Gamma))$ satisfying $a(\zeta) = I$ near W_0, $\operatorname{supp} a \subset W_1$. We define the cutoff operator $X(z)$ as

$$(X(z)f)(x) = \sum_{k=1}^{N_0} \phi_k(x)((s^k)^{-1})^*[Op_z(a^k)(s^k)^*[\phi_k(\cdot)f(\cdot)]](x), \tag{4.1}$$

where $a^k(\sigma, \eta) = (\tilde{s}^k)^* a(\sigma, \eta)$. Here, for $u(\sigma') \in C_0^\infty(\widetilde{U}_k)$, $Op_z(a^k)$ is a pseudodifferential operator with a parameter z defined as

$$Op_z(a^k)u(\sigma) = (2\pi)^{1-n} z^{n-1} \int\int e^{\sqrt{-1}z(\sigma-\sigma')\cdot\eta} a^k(\sigma, \eta) u(\sigma')\, d\sigma'\, d\eta.$$

Note that $Op_z(a^k)$ is well defined for any $z \in \mathbf{C}$ since $\operatorname{supp} a^k$ is compact (for the properties of pseudodifferential operators with a parameter, see [4] or [20]).

Choose functions $\widetilde{\psi}_k(\sigma) \in C_0^\infty(\widetilde{U}_k)$ satisfying $\widetilde{\psi}_k(\sigma) = 1$ near $\operatorname{supp} \widetilde{\phi}_k$, where $\widetilde{\phi}_k(\sigma) = ((s^k)^*\phi_k)(\sigma) \in C_0^\infty(\widetilde{U}_k)$, and cutoff functions $\varphi_k(x) \in C_0^\infty(\mathbf{R}^n)$ such that

$$\begin{aligned} &\varphi_k(x^k(r, \sigma)) = 1 \quad \text{near } r = 0, \sigma \in \operatorname{supp} \widetilde{\phi}_k, \\ &\operatorname{supp} \varphi_k \subset \{x^k(r, \sigma)\ ;\ |r| < b_0, \sigma \in \widetilde{U}_k\}. \end{aligned}$$

Now, we construct an approximation $U_N(z) = \sum_{k=1}^{N_0} U_N^k(z)$ of the Poisson operator $U^{\pm}(z)$ of the form

$$\begin{aligned} U_N^k(z)f(x^k(r, \sigma)) = \varphi_k(x^k(r, \sigma))(2\pi)^{1-n} z^{n-1} \int\int e^{\sqrt{-1}z(\sigma-\sigma')\cdot\eta} u_z^{k,N}(r, \sigma, \eta) \\ \cdot \widetilde{\psi}_k(\sigma')(s^k)^*(\phi_k(\cdot)f(\cdot))(\sigma')\, d\sigma' d\eta \end{aligned}$$

so that $U_N(z)$ satisfies the following equation:

$$\begin{cases} (A(x, \partial_x) + z^2)U_N(z)f(x) = V_N(z)f(x) & \text{in} \quad \Omega, \\ U_N(z)f(x) = X(z)f(x) & \text{on} \quad \Gamma, \end{cases} \tag{4.2}$$

where the operator $V_N(z)$ represents the remainder terms in the approximation.

We seek $u_z^{k,N}(r, \sigma, \eta)$ as $u_z^{k,N}(r, \sigma, \eta) = \sum_{l=0}^{N-1} u_{z,l}^k(r, \sigma, \eta)$. Putting $U_N(z)$ into (4.2), we obtain equations which $u_{z,l}^k$ have to satisfy on each local coordinate. From now on, we do not write the suffix k, referring to the local coordinate U_k, if it is clear from the context.

The change of variables $(r,\sigma)\mapsto x(r,\sigma)=s(\sigma)-r\nu(s(\sigma))$ transforms the operator $A(x,\partial_x)$ to $\widetilde{A}(r,\sigma,\partial_r,\partial_\sigma)$, which yields $\exp(-\sqrt{-1}z\sigma\cdot\eta)\{\widetilde{A}(r,\sigma,\partial_r,\partial_\sigma)+z^2\}(\exp(\sqrt{-1}z\sigma\cdot\eta)v)=\{\widetilde{A}^{(0)}(r,\sigma,\partial_r,\sqrt{-1}z\eta)+\widetilde{A}^{(1)}(r,\sigma,\partial_r,\partial_\sigma,\sqrt{-1}z\,\eta)+\widetilde{A}^{(2)}(r,\sigma,\partial_\sigma)\}v$, where $\widetilde{A}^{(j)}$ is a homogeneous polynomial in ∂_r and $\sqrt{-1}z\eta$ of order $2-j$. Expanding each coefficient of $\widetilde{A}^{(j)}$ into a power series of r at $r=0$, putting the terms of $\exp(-\sqrt{-1}z\,\sigma\cdot\eta)\{\widetilde{A}(r,\sigma,\partial_r,\partial_\sigma)+z^2\}(\exp(\sqrt{-1}z\sigma\cdot\eta)u_z^N(r,\sigma,\eta))$ with the same order together by the rule that the terms ∂_r and z have order 1, and the power $r^j(j\geq 0)$ and the function $u_{z,j}(r,\sigma,\eta)$ have order $-j$, we obtain

$$\text{(4.3)}\qquad \begin{cases} (\widetilde{A}^{(0)}(0,\sigma,\partial_r,\sqrt{-1}z\eta)+z^2)u_{z,l}(r,\sigma,\eta)=\mathcal{R}_l(r,\sigma,\eta;z) \\ \qquad\qquad\qquad \text{in} \quad r>0, \\ u_{z,l}(0,\sigma,\eta)=\delta_{0,l}\widetilde{\phi}(\sigma)a(\sigma,\eta) \end{cases}$$

for $l=0,1,\dots,N-1$, where $\delta_{0,l}$ is Kronecker's delta and

$$\begin{aligned}\mathcal{R}_l(r,\sigma,\eta;z)=-\Big\{&\sum_{j=1}^{l}\frac{1}{j!}r^j(\partial_r^j\widetilde{A}^{(0)})(0,\sigma,\partial_r,\sqrt{-1}z\eta)u_{z,l-j}\\ &+\sum_{j=0}^{l-1}\frac{1}{j!}r^j(\partial_r^j\widetilde{A}^{(1)})(0,\sigma,\partial_r,\partial_\sigma,\sqrt{-1}z\eta)u_{z,l-1-j}\\ &+\sum_{j=0}^{l-2}\frac{1}{j!}r^j(\partial_r^j\widetilde{A}^{(2)})(0,\sigma,\partial_\sigma)u_{z,l-2-j}\Big\}.\end{aligned}$$

Here we have used the convention for the summentions with respect to j in the right-hand side, that is, each of their sums is zero if there is no j in the summations.

To solve (4.3), we follow the argument in Chapter 6 of Kumano-go [13]; however, we need further consideration to handle complex parameter z. Since $\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta)=\sigma_p(A)(s(\sigma),\theta(\sigma,\eta,\zeta))$, the set $\mathcal{N}(\sigma,\eta)=\{\zeta\in\mathbf{C}\,;\det(I-\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))=0\}$ does not intersect the real axis for any $(\sigma,\eta)\in\tilde{s}^{-1}(W_1)\subset\tilde{s}^{-1}(\mathcal{E}_\Gamma)$ (cf. Definition 1.1). Choose a bounded Jordan curve $\mathcal{C}$ enclosing the set $\cup_{(\sigma,\eta)\in\tilde{s}^{-1}(W_1)}\{\zeta\in\mathcal{N}(\sigma,\eta)\,;\operatorname{Im}\zeta>0\}$ and belonging to the set $\{\zeta\,;2d_0<\operatorname{Im}\zeta<2d_0'\}$ for some fixed constants $0<d_0<d_0'$.

Because of the asymptotic behavior

$$-(I-\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))^{-1}=\Big(\sum_{i,l=1}^{n}C_{il}(s(\sigma))\nu_i(s(\sigma))\nu_l(s(\sigma))\Big)^{-1}\zeta^{-2}+O(|\zeta|^{-3})$$
$$\text{as}\quad |\zeta|\to\infty\quad \text{in } \mathbf{C},$$

and the existence of $(I-\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))^{-1}$ for any $\zeta\in\mathbf{R}$, $(\sigma,\eta)\in\tilde{s}^{-1}(W_1)$, $(\int_{\mathcal{C}}(I-\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))^{-1}\,d\zeta)^{-1}$ exists for any $(\sigma,\eta)\in\tilde{s}^{-1}(W_1)$. Thus, we can define the function $v(z,\sigma,\eta)$ by

$$\text{(4.4)}\qquad \begin{aligned}v(z,\sigma,\eta)=\int_{\mathcal{C}}&e^{\sqrt{-1}z\cdot\zeta}(I-\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))^{-1}\,d\zeta\\ &\cdot\Big(\int_{\mathcal{C}}(I-\widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))^{-1}\,d\zeta\Big)^{-1}.\end{aligned}$$

LEMMA 4.1. *There exists a constant $\delta_1 > 0$ depending only upon Γ, W_1, $C_{il}(x)|_\Gamma$, and the solutions $u_{z,l}(r,\sigma,\eta)$ of the equation* (4.3) *such that*

(1) $u_{z,l}$ *are* C^∞ *in* $(r,\sigma,\eta) \in [0,\infty) \times \tilde{s}^{-1}(W_1)$,*holomorphic in* z,$|\mathrm{Im}\ z| \leq \delta_1 \mathrm{Re}\ z$,

(2) $|(\partial_r^{\beta_0}\partial_\sigma^\beta\partial_\eta^\alpha u_{z,l})(r,\sigma,\eta)| \leq C_{\beta_0,\beta,\alpha}|z|^{-l+\beta_0}e^{-d_l(\mathrm{Re}\ z)r}$ *for any* $(r,\sigma,\eta) \in [0,\infty) \times \tilde{s}^{-1}(W_1), |\mathrm{Im}\ z| \leq \delta_1 \mathrm{Re}\ z$, *where* $d_l = 2^{-1}(1+2^{-l})d_0 > 0$,

(3) *there is a function* $\tilde{u}_l(r,\sigma,\eta) \in C^\infty([0,\infty) \times \tilde{s}^{-1}(W_1))$*satisfying* $u_{z,l}(r,\sigma,\eta) = z^{-l}\tilde{u}_l(zr,\sigma,\eta)$ *for any* $(r,\sigma,\eta) \in [0,\infty) \times \tilde{s}^{-1}(W_1), z > 0$.

Proof. For $l = 0$, $u_{z,0}$ is given by $u_{z,0}(r,\sigma,\eta) = \widetilde{\phi}(\sigma)a(\sigma,\eta)v(zr,\sigma,\eta)$. We assume Lemma 4.1 is true for $u_{z,j}(j = 0,1,\dots,l-1)$. Following Kumano-go [13], for $z > 0$, we set

$$w_z(r,\sigma,\eta) = (2\pi)^{-1}z^{-1}\int_{\mathbf{R}} e^{\sqrt{-1}z\zeta r}(I - \widetilde{A}^{(0)}(0,\sigma,\zeta,\eta))^{-1} \cdot \mathcal{F}_z[\widetilde{\mathcal{R}}_l(\cdot,\sigma,\eta;z)](\zeta)\,d\zeta,$$

where $\mathcal{F}_z[k](\zeta) = \int_{\mathbf{R}} \exp\{-\sqrt{-1}z\zeta r'\}k(r')\,dr'$, $\widetilde{\mathcal{R}}_l(r,\sigma,\eta;z) = \mathcal{R}_l(r,\sigma,\eta;z)$ if $r \geq 0$, $\widetilde{\mathcal{R}}_l(r,\sigma,\eta;z) = 0$ if $r < 0$. Then we can give $u_{z,l}$ as $u_{z,l}(r,\sigma,\eta) = w_z(r,\sigma,\eta) - v(zr,\sigma,\eta)w_z(0,\sigma,\eta)$. Thus, (3) in Lemma 4.1 follows from assumption of induction, since $\mathcal{F}_z[\widetilde{\mathcal{R}}_l(\cdot,\sigma,\eta;z)](\zeta) = z^{-1}\mathcal{F}_1[\widetilde{\mathcal{R}}_l(z^{-1}\cdot,\sigma,\eta;z)](\zeta)$ holds.

We continue w_z analytically in z. Change of variable and change of contour imply

$$(\partial_r^j w_z)(r,\sigma,\eta) = (2\pi)^{-1}z^{-2}\int_{-\infty+\sqrt{-1}\gamma}^{\infty+\sqrt{-1}\gamma}(\sqrt{-1}\zeta)^j e^{\sqrt{-1}\zeta r} \cdot(I - \widetilde{A}^{(0)}(0,\sigma,z^{-1}\zeta,\eta))^{-1}\mathcal{F}_1[\widetilde{\mathcal{R}}_l(\cdot,\sigma,\eta;z)](\zeta)\,d\zeta \tag{4.5}$$

for any $j = 0,1$, $0 \leq \gamma \leq d_l'z$, $z > 0$, where $d_l' = 2^{-1}(d_l + d_{l-1})$. Since the properties (1), (2) for $u_{z,j}$ $(j = 0,1,\dots,l-1)$ ensure holomorphicity of $\mathcal{F}_1[\widetilde{\mathcal{R}}_l(\cdot,\sigma,\eta;z)](\zeta)$ in $\mathrm{Im}\ \zeta \leq d_l'\mathrm{Re}\ z$ with an estimate

$$|\zeta^j\partial_\sigma^\beta\partial_\eta^\alpha(\mathcal{F}_1[\widetilde{\mathcal{R}}_l(\cdot,\sigma,\eta;z)](\zeta))| \leq C_{\alpha,\beta}|z|^{-l+2+j} \tag{4.6}$$

for any $\zeta, z \in C$, $\mathrm{Im}\ \zeta \leq d_l'\mathrm{Re}\ z$, $(\sigma,\eta) \in \tilde{s}^{-1}(W_1)$, $|\mathrm{Im}\ z| \leq \delta_1|\mathrm{Re}\ z|$, and $j = 0,1$.

Now we move the parameter z in the complex plane and seek the region of z in which the integral (4.5) is still valid. We set $\zeta' = z^{-1}\zeta$. Since $\mathrm{Im}\ \zeta' = -\frac{\mathrm{Im}\ z}{\mathrm{Re}\ z}(\mathrm{Re}\ \zeta') + (\frac{\mathrm{Re}\ z}{|z|^2} + \frac{(\mathrm{Im}\ z)^2}{|z|^2\mathrm{Re}\ z})\mathrm{Im}\ \zeta$, the image of the line $\mathrm{Im}\ \zeta = \gamma$ by the map $\zeta' = z^{-1}\zeta$ is contained in the region $|\mathrm{Im}\ \zeta'| \leq \delta'|\mathrm{Re}\ \zeta'| + d_l'$ if z stays in $|\mathrm{Im}\ z| \leq \delta'|\mathrm{Re}\ z|$, and $0 \leq \gamma \leq d_l'\mathrm{Re}\ z$. Choose $\delta_2 > 0$ depending only upon Γ, W_1, and $C_{il}(x)|_\Gamma$ $(i,l = 1,\dots,n)$ as $\{\zeta'\,;\,|\mathrm{Im}\ \zeta'| \leq \delta_2|\mathrm{Re}\ \zeta'| + d_l'\,\} \cap \mathcal{N}(\sigma,\eta) = \phi$ for any $(\sigma,\eta) \in \tilde{s}^{-1}(W_1)$ so that the integral (4.5) is well defined in $|\mathrm{Im}\ z| \leq \min\{\delta_1,\delta_2\}\mathrm{Re}\ z$, $0 < \gamma \leq \tilde{d}_l\mathrm{Re}\ z$, and $\partial_r^j w_z$ $(j = 0,1)$ is holomorphic in z, where $\delta_1 > 0$ is the constant specified in Lemma 4.1 for $u_{z,j}$ $(j = 0,1,\dots,l-1)$. Thus, $\partial_r^j u_{z,l}$ $(j = 0,1)$ can be also continued analytically in $|\mathrm{Im}\ z| \leq \min\{\delta_1,\delta_2\}\mathrm{Re}\ z$. From the equation of $u_{z,l}$ for $z > 0$, it follows that $u_{z,l}$ satisfies (4.3) for z in $|\mathrm{Im}\ z| \leq \min\{\delta_1,\delta_2\}\mathrm{Re}\ z$ and the property (1) in Lemma 4.1 by the analytic continuation.

To show property (2), we choose $\gamma = d_l'\mathrm{Re}\ z$ in the integral (4.5) and change the variable $\mathrm{Re}\ \zeta$ to $\gamma\mathrm{Re}\ \zeta$. Since $1 \geq |z^{-1}\mathrm{Re}\ z| \geq (1+\delta'^2)^{-1/2} > 0$ in $|\mathrm{Im}\ z| \leq \delta'\mathrm{Re}\ z$, by (4.6), we have the estimate of $\partial_r^j\partial_\sigma^\beta\partial_\eta^\alpha u_{z,l}$ for $j = 0,1$ as in property (2). Thus, (4.3) implies the property (2) inductively. By the definition $\delta_2 > 0$ in each step of induction, we can choose $\delta_1 > 0$ in Lemma 4.1 independent of $u_{z,l}$. This completes the proof of Lemma 4.1.

From Lemma 4.1, it follows that

$$\begin{aligned}(4.7)\qquad &|\partial_r^{\beta_0}\partial_\sigma^{\beta}\partial_\eta^{\alpha}\{e^{-\sqrt{-1}z\sigma\eta}(\widetilde{A}(r,\sigma,\partial_r,\partial_\sigma)+z^2)(e^{\sqrt{-1}z\sigma\eta}u_z^N(r,\sigma,\eta))\}|\\ &\qquad\le C_{\beta_0,\beta,\alpha,N}|z|^{-N+2+\beta_0}e^{-(d_0/2)(\mathrm{Re}\ z)r}\end{aligned}$$

for any $(r,\sigma,\eta)\in[0,\infty)\times\tilde{s}^{-1}(W_1)$, $|\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z$ and $u_z^N(0,\sigma,\eta)=\widetilde{\phi}(\sigma)a(\sigma,\eta)$. By the procedure of the construction, there is a constant $a>0$ with $\Gamma\subset B_a$ satisfying

$$(4.8)\qquad \mathrm{supp}\ (V_N(z)f)\subset\overline{\Omega}\cap B_a\quad\text{for any } f\in C^\infty(\Gamma)\text{ and } |\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z.$$

Furthermore, from (4.7) it follows that for any $b_0>0$, there exist constants $C_{N,b_0}>0$, $b_1>0$ such that

$$\begin{aligned}(4.9)\qquad &\|V_N(z)f\|_{L^2(\Omega)}\le C_{N,b_0}|z|^{-N+2}\|f\|_{L^2(\Gamma)}\\ &\text{for any } f\in C^\infty(\Gamma),\ |\mathrm{Im}\ z|\le b_0\log(\mathrm{Re}\ z),\ \mathrm{Re}\ z\ge b_1.\end{aligned}$$

In fact, it follows from the L^2-boundness theorem of pseudodifferential operator (cf. Gérard [4]), since supp $u_{z,l}^k\subset[0,\infty)\times((\tilde{s}^k)^{-1}(W_1)\cap(\text{ supp }\widetilde{\phi}_k\times\mathbf{R}^n))$ for $l=0,1,\dots,N-1$, and we can consider that the operator $V_N(z)$ is a finite sum of pseudodifferential operators on $\widetilde{U}_k$ with real parameter Re z having symbols estimated like (4.7) as long as z stays in $|\mathrm{Im}\ z|\le b_0\log(\mathrm{Re}\ z)$ and $|\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z$.

Using the approximation $U_N(z)$ of $U^\pm(z)$, we define the approximation $B_N(z)$ of the Neumann operator by

$$B_N(z)f(x)=(N(x,\partial_x)U_N(z)f)|_\Gamma.$$

By (4.2), the uniqueness of problem (3.1) implies

$$\begin{aligned}&U^\pm(z)X(z)=U_N(z)-R^\pm(z)V_N(z)+U^\pm(z)\gamma_\Gamma\cdot R^\pm(z)V_N(z)\\ &\qquad\text{for any }\ \pm\,\mathrm{Im}\ z<0,|\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z,\end{aligned}$$

where γ_Γ is the trace operator on Γ. From the definition of $T^\pm(z)$, it follows that

$$T^\pm(z)X(z)=B_N(z)+T^\pm(z)\gamma_\Gamma R^\pm(z)V_N(z),$$

which yields

$$\begin{aligned}(4.10)\qquad &X(z)=(T^\pm(z))^{-1}B_N(z)+\gamma_\Gamma R^\pm(z)V_N(z)\\ &\text{for any }\ z\in\widetilde{\mathbf{C}}_\pm,\pm\mathrm{Im}\ z<0,|\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z,\mathrm{Re}\ z\ge 1.\end{aligned}$$

Note that (4.10) is still valid in a region in $|\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z$ wherever $(T^\pm(z))^{-1}$ and $R^\pm(z)$ are continued analytically because of the analyticity of $V_N(z)$ in $|\mathrm{Im}\ z|\le\delta_1\mathrm{Re}\ z$.

From the form of $U_N(z)$, the operator $B_N(z)$ is of the form

$$\begin{aligned}(4.11)\qquad B_N(z)f(x)=\sum_{k=1}^{N_0}&\varphi_k(x)((s^k)^{-1})^*[Op_z(b_z^{k,N}(\sigma,\eta)\widetilde{\psi}_k(\sigma'))\\ &\cdot(s^k)^*(\phi_k(\cdot)f(\cdot))](x),\end{aligned}$$

where $b_z^{k,N}(\sigma,\eta)=\sum_{j=0}^{N-1}z^{1-j}b_j^k(\sigma,\eta)$ with $b_j^k\in C_0^\infty((\tilde{s}^k)^{-1}(W_1))$ independent of N. Moreover, there exists a function $l_0(\zeta)\in C^\infty(\mathcal{E}_\Gamma)$ such that $b_0^k(\sigma,\eta)=\widetilde{\phi}_k(\sigma)a^k(\sigma,\eta)$ $((\tilde{s}^k)^*l_0)(\sigma,\eta)$. Note that $((\tilde{s}^k)^*l_0)(\sigma,\eta)$ is given by

$$\begin{aligned}((\tilde{s}^k)^*l_0)(\sigma,\eta)=&-\sum_{i,l=1}^{n}C_{il}(s^k(\sigma))\nu_i(s^k(\sigma))\nu_l(s^k(\sigma))(\partial_r v^k)(0,\sigma,\eta)\\&+\sqrt{-1}\sum_{i,l=1}^{n}C_{il}(s^k(\sigma))\nu_i(s^k(\sigma))\theta_l^k(\sigma,\eta,0),\end{aligned}$$

where $v^k(r,\sigma,\eta)$ is defined as (4.4) in the coordinate $[0,\infty)\times\widetilde{U}_k$. In what follows, we call the function l_0 the principal part of the operator $B_N(z)$.

Before finishing this section, we describe some properties of the operators $B_N(z)$, $X(z)$ with $\mu=\operatorname{Im}z$ as a family of pseudodifferential operators with real parameter $\lambda=\operatorname{Re}z$. We set $\Lambda_{a_0,a_1}=\{z=\lambda+\sqrt{-1}\mu\in\mathbf{C}|\,|\mu|\le a_0\log\lambda,\lambda\ge a_1\}$. Denote by $\Psi_{\rho,\delta}^{m,k}(\Gamma:[\lambda_1,\infty))$ the space of pseudodifferential operators with real parameter on Γ with local symbols $a_\lambda(\sigma,\eta)$ satisfying

$$|\partial_\sigma^\beta\partial_\eta^\alpha a_\lambda(\sigma,\eta)|\le C_{\alpha,\beta}|\lambda|^{k+\rho|\alpha|+\delta|\beta|}(1+|\eta|)^{m-|\alpha|}\qquad\text{for all }\sigma,\eta\in\mathbf{R}^n\text{ and }\lambda\ge\lambda_1.$$

PROPOSITION 4.2. (1) *For any fixed $a_0>0$, the operator $B_N(\lambda+\sqrt{-1}\mu)$ belongs to $\Psi_{0,0}^{-\infty,1}(\Gamma:[1,\infty))$ uniformly in μ with $\lambda+\sqrt{-1}\mu\in\Lambda_{a_0,1}$. The principal symbol $\sigma_p(B_N(\lambda+\sqrt{-1}\mu))$ is*

$$\sigma_p(B_N(\lambda+\sqrt{-1}\mu))=z\Big\{a\cdot l_0-\sqrt{-1}\frac{\mu}{\lambda}H_{rad}(a\cdot l_0)\Big\},$$

where H_{rad} is the radial vector field on $T^*(\Gamma)$ defined by $(\tilde{s}^{-1})_*H_{rad}=\sum_{j=1}^n\eta_j\frac{\partial}{\partial\eta_j}$ for any local triviality $\tilde{s}$.

(2) *The formal adjoint operator $(B_N(\lambda+\sqrt{-1}\mu))^*$ belongs to $\Psi_{0,0}^{-\infty,1}(\Gamma:[1,\infty))$ uniformly in μ with $\lambda+\sqrt{-1}\mu\in\Lambda_{a_0,1}$ and it satisfies*

$$\sigma_p((B_N(\lambda+\sqrt{-1}\mu))^*)=\overline{z}\Big\{a\cdot l_0+\sqrt{-1}\frac{\mu}{\lambda}H_{rad}(a\cdot l_0)\Big\}.$$

In particular, $\sigma_p((B_N(\lambda+\sqrt{-1}\mu))^*)-\sigma_p(B_N(\lambda-\sqrt{-1}\mu))\in\Psi_{0,0}^{-\infty,0}(\Gamma:[1,\infty))$.

(3) *$X(\lambda+\sqrt{-1}\mu)\in\Psi_{0,0}^{-\infty,1}(\Gamma:[1,\infty))$ uniformly in μ with $\lambda+\sqrt{-1}\mu\in\Lambda_{a_0,1}$ and we have*

$$\sigma_p(X(\lambda+\sqrt{-1}\mu))=a-\sqrt{-1}\frac{\mu}{\lambda}H_{rad}(a).$$

5. Properties of the principal part $l_0(\zeta)$. By the definition of the surface impedance tensor $Z(\zeta,\tau)$ and the principal part $l_0(\zeta)$ of the approximation operator $B_N(z)$, we have

$$l_0(\zeta)=\|\zeta\|_\Gamma Z(\zeta,1)\qquad\text{for any }\zeta\in\mathcal{E}_\Gamma. \tag{5.1}$$

In this section, we show properties of $l_0(\zeta)$, which are crucial in proving Theorems 0.1 and 0.3 by the surface impedance tensor.

We set $\Sigma = \{\zeta \in \mathcal{E}_\Gamma\,;\, \det(Z(\zeta,1)) = 0\}$. The set Σ is a closed subset in $\mathcal{E}_\Gamma$.

LEMMA 5.1. *There exist open sets W_0 and W_1 in $\mathcal{E}_\Gamma$ such that $\Sigma \subset W_0$, $\overline{W_0} \subset W_1$, $\overline{W_1} \subset \mathcal{E}_\Gamma$, where the closure $\overline{W_i}$ of W_i $(i = 0, 1)$ is taken in the topology of $T^*(\Gamma)$. Furthermore,*

$$l_0(\zeta) \in C^\infty(\overline{W_1}) \textit{ is a Hermite matrix for any } \zeta \in \overline{W_1};$$
$$l_0(\zeta) \textit{ is an invertible matrix for any } \zeta \in W_0 \setminus \Sigma.$$

Note. In what follows, we use the open sets W_0 and W_1 chosen in Lemma 5.1 as the open sets used for constructing the approximation operator $B_N(z)$ in section 4.

Proof. Take a local coordinate $s(\sigma)$ in section 1. The surface impedance tensor $Z(\sigma,\eta,\tau)$ expressed by the local triviality $\tilde{s}$ induced by s has eigenvalues $\zeta_1(\sigma,\eta,\tau), \dots, \zeta_n(\sigma,\eta,\tau)$ which are real-valued, continuous in $(\sigma,\eta) \in T^*(\widetilde{U})$, $0 < \tau \le \hat{v}(\sigma,\eta)|\eta|_\Gamma$ and real analytic in $0 < \tau < \hat{v}(\sigma,\eta)|\eta|_\Gamma$, and homogeneous of order 0 in (η,τ), where $\hat{v}(\sigma,\eta) = (\tilde{s}^* v_L)(\sigma,\eta)$. Moreover, for any $j = 1, 2, \dots, n$, we have

$$\partial_\tau \zeta_j(\sigma,\eta,\tau) < 0 \quad \text{in } 0 < \tau < \hat{v}(\sigma,\eta)|\eta|_\Gamma, \tag{5.2}$$
$$\zeta_j(\sigma,\eta,0) > 0 \tag{5.3}$$

(cf. Lemma 4.1 in Nakamura [15]).

Set $I = \{j\,;\, \zeta_j(\sigma,\eta,\hat{v}(\sigma,\eta)|\eta|_\Gamma) < 0$ for some $(\sigma,\eta) \in T^*(\widetilde{U})\}$. Since property (5.2) implies $\tilde{s}^{-1}(\Sigma) = \cup_{j\in I}\{(\sigma,\eta) \in T^*(\widetilde{U})\,;\, \zeta_j(\sigma,\eta,1) = 0, 1 < \hat{v}(\sigma,\eta)|\eta|_\Gamma\}$, by (5.2) and (5.3) we have $\tilde{s}^{-1}(\Sigma) \subset \{(\sigma,\eta) \in T^*(\widetilde{U})\,;\, 1 + \epsilon_0 \le \hat{v}(\sigma,\eta)|\eta|_\Gamma \le \epsilon_0^{-1}\} \subset \mathcal{E}_\Gamma$ with a fixed $\epsilon_0 > 0$. This means the set Σ is a compact set in $\mathcal{E}_\Gamma$ so that we can choose open sets W_0 and W_1 as in Lemma 5.1. Since $Z(\zeta,\tau)$ is a Hermite matrix, the rest of the proof is obvious.

Next we show how the condition (ERW) reflects on a property of $l_0(\zeta)$, which is given by Nakamura [15] in the three dimensional case.

PROPOSITION 5.2. *If we assume that the condition* (ERW) *holds, there exists a local coordinate $s : \widetilde{U} \to U$ and a point $(\sigma^0,\eta^0) \in \tilde{s}^{-1}(\mathcal{E}_\Gamma)$, an open neighborhood $\widetilde{W}$ of (σ^0,η^0) in $\tilde{s}^{-1}(W_0)$, a real-valued C^∞-function $\lambda(\sigma,\eta)$ on $\widetilde{W}$, and an $n \times n$-matrix valued C^∞-function $q(\sigma,\eta)$ on $\widetilde{W}$ such that*

$$\lambda(\sigma^0,\eta^0) = 1,\ (\nabla_\eta\lambda)(\sigma^0,\eta^0) \ne 0,\ q(\sigma^0,\eta^0) \ne 0 \ \textit{and}$$
$$(\tilde{s}^* l_0)(\sigma,\eta)q(\sigma,\eta) = 0 \ \textit{on } \lambda(\sigma,\eta) = 1,\ (\sigma,\eta) \in \widetilde{W}.$$

Proof. From (5.1), it suffices to show the same statement for the function $Z(\zeta,1)$. By the condition (ERW), we can choose a local coordinate $\widetilde{U} \to U \subset \Gamma$ such that $\zeta_1(\sigma',\eta',\hat{v}(\sigma',\eta')|\eta'|_\Gamma) < 0$ for some $(\sigma',\eta') \in T^*(\widetilde{U})$ by reenumeration of the eigenvalues $\zeta_1, \zeta_2, \dots, \zeta_n$ if it is necessary. From Lemma 4.1 of Nakamura [15], we can find a conic neighborhood $\widetilde{W}_1$ of (σ',η') in $T^*(\widetilde{U})$, a real-valued, positive, and continuous function $\lambda(\sigma,\eta)$ on $\widetilde{W}_1$ homogeneous of order 1 in η satisfying

$$\zeta_1(\sigma,\eta,\lambda(\sigma,\eta)) = 0, \quad 0 < \lambda(\sigma,\eta) < \hat{v}(\sigma,\eta)|\eta|_\Gamma \quad \text{for any } (\sigma,\eta) \in \widetilde{W}_1.$$

We can reenumerate the eigenvalues $\zeta_2, \dots, \zeta_n$ satisfying $\zeta_j(\sigma', \eta', \lambda(\sigma', \eta')) = 0$ for $j = 1, 2, \dots, r_0$ and $\zeta_{j'}(\sigma', \eta', \lambda(\sigma', \eta')) \neq 0$ for $j' = r_0 + 1, \dots, n$ with an integer $1 \leq r_0 \leq n$. We can assume that $\zeta_j(\sigma, \eta, \lambda(\sigma, \eta)) \neq 0$ for any $j = r_0 + 1, \dots, n$, $(\sigma, \eta) \in \widetilde{W}_1$. Starting from $\widetilde{W}_1$, define sequences of open sets $\widetilde{V}_j$ and $\widetilde{W}_j$ $(j = 2, \dots, r_0)$ as $\widetilde{V}_j = \{(\sigma, \eta) \in \widetilde{W}_{j-1}; \zeta_j(\sigma, \eta, \lambda(\sigma, \eta)) \neq 0\}$ and $\widetilde{W}_j = \widetilde{V}_j$ if $\widetilde{V}_j \neq \phi$, $\widetilde{W}_j = \widetilde{W}_{j-1}$ if $\widetilde{V}_j = \phi$. Set $\tilde{I} = \{1\} \cup \{j;\ \widetilde{V}_j = \phi\}$. Then it is obvious that the open set $\widetilde{W}_{r_0}$ is not empty and $\zeta_j(\sigma, \eta, \lambda(\sigma, \eta)) = 0$, $\zeta_{j'}(\sigma, \eta, \lambda(\sigma, \eta)) \neq 0$ for any $(\sigma, \eta) \in \widetilde{W}_{r_0}$, $j \in \tilde{I}$, $j' \in \{1, \dots, n\} \setminus \tilde{I}$.

Now, we show $\lambda(\sigma, \eta) \in C^\infty(\widetilde{W}_{r_0})$. Set $F(\sigma, \eta, \tau) = \det(Z(\sigma, \eta, \tau))$. Since $F(\sigma, \eta, \tau) = \Pi_{j=1}^n \zeta_j(\sigma, \eta, \tau)$ for any $(\sigma, \eta) \in \widetilde{W}_{r_0}$, from (5.2) we have

$$(\partial_\tau^l F)(\sigma, \eta, \lambda(\sigma, \eta)) = 0, \quad (l = 0, 1, \dots, r-1), \quad (\partial_\tau^r F)(\sigma, \eta, \lambda(\sigma, \eta)) \neq 0,$$

where we denote by r the number of the element of $\tilde{I}$. This gives $\lambda(\sigma, \eta) \in C^\infty(\widetilde{W}_{r_0})$ since $\lambda(\sigma, \eta)$ is an implicit function of equation $(\partial_\tau^{r-1} F)(\sigma, \eta,\ \lambda(\sigma, \eta)) = 0$ with $\partial_\tau(\partial_\tau^{r-1} F)(\sigma, \eta, \lambda(\sigma, \eta)) \neq 0$.

We can choose a point $(\sigma^0, \eta^0) \in \widetilde{W}_{r_0}$ and an open neighborhood $\widetilde{W}$ satisfying $\lambda(\sigma^0, \eta^0) = 1$, $\widetilde{W} \subset \widetilde{W}_{r_0} \cap \mathcal{E}_\Gamma$, because (5.3) ensures $0 < \lambda(\sigma, \eta) < \hat{v}(\sigma, \eta)|\eta|_\Gamma$ in $\widetilde{W}_{r_0}$. Since $\lambda(\sigma, \eta)$ is positive and homogeneous of order 1, we have $(\nabla_\eta \lambda)(\sigma, \eta) \neq 0$ on $\widetilde{W}$. We can also take a constant $c_0 > 0$ satisfying $|\zeta_j(\sigma, \eta, 1)| < c_0$, $|\zeta_l(\sigma, \eta, 1)| > c_0$ on $\widetilde{W}$ for $j \in \tilde{I}$, $l \in \{1, \cdots, n\} \setminus \tilde{I}$ by shrinking $\widetilde{W}$ if it is necessary. Set $q(\sigma, \tau) = (2\pi\sqrt{-1})^{-1} \int_{|\zeta| = c_0} (\zeta - Z(\sigma, \eta, 1))^{-1}\, d\zeta \in C^\infty(\widetilde{W})$. Then $Z(\sigma, \eta, 1) q(\sigma, \eta) = 0$ for $(\sigma, \eta) \in \widetilde{W}$ with $\lambda(\sigma, \eta) = 1$ is obvious because $\zeta_j(\sigma, \eta, 1) = 0$ on $\lambda(\sigma, \eta) = 1$, $(\sigma, \eta) \in \widetilde{W}$ $j \in \tilde{I}$. Since $q(\sigma^0, \eta^0)$ is the eigenprojection of $Z(\sigma^0, \eta^0, 1)$ to the eigenspace $\mathrm{Ker}, (Z(\sigma^0, \eta^0, 1)) \neq \{0\}$, $q(\sigma^0, \eta^0) \neq 0$. This completes the proof of Proposition 5.2.

6. Asymptotic null solution of $B_N(z)$.

Theorems 0.1 and 0.3 are proved by using asymptotic null solutions of $B_N(z)$ described in the following theorem.

Theorem 6.1. *Assume that the condition* (ERW) *holds. Then, for any integer N, there exist a sequence of function f_j $(j = 1, 2, \dots)$ and a sequence z_j $(j = 1, 2, \dots)$, a constant $C_N > 0$ such that the following* (1) $\sim$ (3) *hold. These are*

(1) $f_j \in C^\infty(\Gamma)$, $\quad \|f_j\|_{L^2(\Gamma)} = 1$ *for any* $j = 1, 2, \dots$,

(2) *for any integer* $N' > 0$, *there is a constant* $C_{N,N'} > 0$ *such that*

$$\|(I - X(z_j)) f_j\|_{L^2(\Gamma)} \leq C_{N,N'} |\mathrm{Re}\ z_j|^{-N'}$$

$$\|B_{2N}(z_j) f_j\|_{L^2(\Gamma)} \leq C_{N,N'} |\mathrm{Re}\ z_j|^{-N'} \quad \textit{for any } j = 1, 2, \dots ,$$

(3) $|\mathrm{Im}\ z_j| \leq C_N |\mathrm{Re}\ z_j|^{-2N+1}$ *for any* $j = 1, 2, \dots$.

Proof of Theorems 0.1 *and* 0.3. If Theorem 0.1 or Theorem 0.3 is not true, from the equality (4.10), the estimate (4.9), the property (4.8), and Lemma 3.1, it follows that

$$\begin{aligned} \|X(z) f\|_{L^2(\Gamma)} \leq C|z|^{m_0 + 3n + 7} \{ &\|B_{2N}(z) f\|_{L^2(\Gamma)} \\ &+ |z|^{-2N+2} \|f\|_{L^2(\Gamma)} \} \\ \text{for any } f \in C^\infty(\Gamma), z \in \widetilde{\mathbf{C}}_\pm, &\mathrm{Re}\ z \geq C_1, \\ &|\mathrm{Im}\ z| \leq C_0 |\mathrm{Re}\ z|^{-(m_0 + 3n + 4)}, \end{aligned} \tag{6.1}$$

with some fixed C, C_0, C_1, and $m_0 > 0$. This estimate is inconsistent with Theorem 6.1 for $2N > m_0 + 3n + 9$. Thus, we have proved Theorems 0.1 and 0.3 are true.

The rest of this paper is devoted to showing Theorem 6.1. Hereafter, we fix $a_0 > 0$ and $N \in \mathbf{N}$. Define a pseudodifferential operator $P(z)$ by

$$P(z) = z^{2N-1} B_{2N}(z) + \sqrt{-1}(I - X(z))(z^2 - \triangle_\Gamma)^N \quad z \in \Lambda_{a_0,1},$$

where $-\triangle_\Gamma$ is the Laplace–Beltrami operator on Γ. The expression (4.11) implies that $P(z)$ is a $B(H^{s+2N}(\Gamma), H^s(\Gamma))$-valued entire function for any $s \in \mathbf{R}$. The statement (1) in Proposition 4.2 says $P(\lambda + \sqrt{-1}\mu) \in \Psi_{0,0}^{2N,2N}(\Gamma : [1,\infty))$ uniformly in μ with $\lambda + \sqrt{-1}\mu \in \Lambda_{a_0,1}$. By $\sigma_p(P(\lambda+\sqrt{-1}\mu))(\zeta) = z^{2N-1}\sigma_p(B_{2N}(\lambda+\sqrt{-1}\mu))(\zeta) + \sqrt{-1}(1 - \sigma_p(X(\lambda + \sqrt{-1}\mu)))z^{2N}(1 + \|\zeta\|_\Gamma^2)^N$ and Lemma 5.1, there is a constant $a_1 > 0$ such that $P(\lambda + \sqrt{-1}\mu) \in \Psi_{0,0}^{2N,2N}(\Gamma : [1,\infty))$ $(\lambda + \sqrt{-1}\mu \in \Lambda_{a_0,a_1})$ is elliptic in $T^*(\Gamma) \setminus \Sigma$ including every infinite point (cf. [21]; for definition of ellipticity for pseudodifferential operator with parameter, see [4]).

We explain our plan to show Theorem 6.1. Consider $P(z)$ as an operator on $L^2(\Gamma)$ with domain $D(P(z)) = H^{2N}(\Gamma)$ for any $z \in \Lambda_{a_0,a_1}$.

Proposition 6.2. *If we assume the condition* (ERW) *holds, there exists a sequence* $z_j \in \Lambda_{a_0,a_1} (j = 1, 2, \dots)$ *such that* $\lim_{j\to\infty} \operatorname{Re} z_j = \infty$ *and* $\operatorname{Ker} P(z_j) \neq \{0\}$ *for any* $j = 1, 2, \dots$.

After proving Proposition 6.2, we choose $f_j \in \operatorname{Ker} P(z_j)$ $\|f_j\|_{L^2(\Gamma)} = 1$ and show that $\{z_j\}$ and $\{f_j\}$ satisfy all properties in Theorem 6.1.

We give a proof of Proposition 6.2 in section 7. In the rest of this section, we show Theorem 6.1 by Proposition 6.2.

Proof of Theorem 6.1. The property (1) is obvious, for $P(\lambda + \sqrt{-1}\mu)$ is elliptic at every infinite point (see [4] for the terminology "infinite point"). Unfortunately, since our pseudodifferential operators with nonreal parameters do not generate algebra, we have to deal with them as pseudodifferential operators with real parameters. Recalling (4.1), we consider $X(\lambda + \sqrt{-1}\mu)$ as a pseudodifferential operator with real parameter λ, that is,

$$X(\lambda + \sqrt{-1}\mu)f(x) = \sum_{k=1}^{N_0} \phi_k(x)((s^k)^{-1})^*[Op_\lambda(\tilde{a}^k_{\lambda,\mu})(s^k)^*[\phi_k(\cdot)f(\cdot)]](x),$$

where for any positive integer M,

$$\begin{aligned}\tilde{a}^k_{\lambda,\mu}(\sigma,\eta) &= \left(1 + \sqrt{-1}\frac{\mu}{\lambda}\right)^n \sum_{|\alpha|<M} \sum_{\beta+\gamma\le\alpha} \frac{\alpha!}{(\alpha-\beta)!\beta!\gamma!(\alpha-\beta-\gamma)!} \\ &\cdot \left(-\sqrt{-1}\frac{\mu}{\lambda}\right)^{|\beta+\gamma|} \lambda^{-|\alpha|+|\beta+\gamma|}\eta^\gamma(-\sqrt{-1}\partial_{\sigma'})^{\alpha-(\beta+\gamma)}\partial_\eta^{\alpha-\beta}(a^k(\sigma,\eta)\widetilde{\psi}_k(\sigma'))|_{\sigma'=\sigma} \\ &+ r_{M,\lambda,\mu}(\sigma,\eta),\end{aligned}$$

with $r_{M,\lambda,\mu}(\sigma,\eta) \in S_{0,0}^{-M,-M+\epsilon}(\mathbf{R}^{n-1} \times \mathbf{R}^{n-1} : [1,\infty))$, $\lambda + \sqrt{-1}\mu \in \Lambda_{a_0,1}$ for any $\epsilon > 0$ and $\widetilde{\psi}_k \in C_0^\infty(\widetilde{U}_k)$ is the one chosen in section 4.

From the estimate $|(1 + \sqrt{-1}\mu/\lambda)^n \sum_{|\alpha|<M}(-\sqrt{-1}\mu/\lambda)^{|\alpha|} - 1| \le C_M|\mu/\lambda|^M$ for any $\lambda + \sqrt{-1}\mu \in \Lambda_{a_0,1}$, we can divide $\tilde{a}^k_{\lambda,\mu}(\sigma,\eta)$ as

$$\tilde{a}^k_{\lambda,\mu}(\sigma,\eta) = \widetilde{\psi}_k(\sigma)a^k(\sigma,\eta) + \tilde{a}^{k,1}_{M,\lambda,\mu}(\sigma,\eta) + \tilde{a}^{k,2}_{M,\lambda,\mu}(\sigma,\eta),$$

where $\tilde{a}^{k,1}_{M,\lambda,\mu}(\sigma,\eta) \in S_{0,0}^{0,0}(\mathbf{R}^{n-1} \times \mathbf{R}^{n-1} : [1,\infty))$, $\operatorname{supp} \tilde{a}^{k,1}_{M,\lambda,\mu}(\sigma,\eta) \subset \widetilde{U}_k \times \mathbf{R}^{n-1}_\eta \setminus (\tilde{s}^k)^{-1}(W_0)$, $\tilde{a}^{k,2}_{M,\lambda,\mu}(\sigma,\eta) \in S_{0,0}^{0,-M+\epsilon}(\mathbf{R}^{n-1}\times\mathbf{R}^{n-1} : [1,\infty))$ for $\epsilon > 0$. Since $(s^k)^*\phi_k(\sigma)\cdot$

$\widetilde{\psi}_k(\sigma) = 1$ on $\widetilde{U}_k$, $a^k(\sigma,\eta) = (\tilde{s}^k)^* a(\sigma,\eta) \in S^{0,0}_{0,0}(\mathbf{R}^{n-1} \times \mathbf{R}^{n-1} : [1,\infty))$, $a = I$ near W_0, the decomposition means

$$I - X(\lambda + \sqrt{-1}\mu) = A^{(1)}_{M,\lambda,\mu} + A^{(2)}_{M,\lambda,\mu},$$

where $A^{(1)}_{M,\lambda,\mu} \in \Psi^{0,0}_{0,0}(\Gamma : [1,\infty))$, $A^{(2)}_{M,\lambda,\mu} \in \Psi^{0,-M+\epsilon}_{0,0}(\Gamma : [1,\infty))$ for any $\epsilon > 0$ and the essential support of the local symbol of $A^{(1)}_{M,\lambda,\mu}$ in any local coordinate does not meet the pullback of W_0 by the local triviality. Hence, for any positive integer M and M', we have

$$\begin{aligned}\left\|(I - X(z_j))(z_j^2 - \triangle_\Gamma)^l f_j\right\|_{L^2(\Gamma)} &\le \sum_{k=1}^{2} \left\|A^{(k)}_{M,\mathrm{Re}\ z_j,\mathrm{Im}\ z_j}(z_j^2 - \triangle_\Gamma)^l f_j\right\|_{L^2(\Gamma)} \\ &\le C_{M',M,l}|\mathrm{Re}\ z_j|^{-M'} + C_M|\mathrm{Re}\ z_j|^{-M+1}\left\|(z_j^2 - \triangle_\Gamma)^l f_j\right\|_{L^2(\Gamma)}.\end{aligned}$$

In fact, for the term $A^{(1)}_{M,\lambda,\mu}((\lambda + \sqrt{-1}\mu)^2 - \triangle_\Gamma)^l$, we can use a standard argument since $P(z_j)f_j = 0$ and $P(\lambda + \sqrt{-1}\mu)$ is elliptic on $T^*(\Gamma) \setminus W_0$ containing every infinite point. For the second term, we use only L^2-boundness theorem for $A^{(2)}_{M,\lambda,\mu}$. Noting that $\|f_j\|_{H^{2N}(\Gamma)} \le C|\mathrm{Re}\ z_j|^{2N}$ for any $j = 1, 2, \dots$, with some fixed constant $C > 0$ which follows from an a priori estimate (7.3) stated in section 7, we obtain $\left\|(I - X(z_j))(z_j^2 - \triangle_\Gamma)^l f_j\right\|_{L^2(\Gamma)} \le C_{N'}|\mathrm{Re}\ z_j|^{-N'}$ for $l = 0, 1, \dots, 2N$. By the definition of $P(z)$, the estimate of $B_{2N}(z_j)f_j$ is obvious. Thus, we have the statement (2) in Theorem 6.1.

Last, we show property (3) in Theorem 6.1. We can assume Im $z_j \ne 0$. From the equality (4.10), the estimate (4.9), and Lemma 3.2, it follows that

$$\begin{aligned}\|X(z)f\|_{L^2(\Gamma)} \le C|\mathrm{Im}\ z|^{-1}\Big\{ &\|B_{2N}(z)f\|_{L^2(\Gamma)} \\ &+ |\mathrm{Re}\ z|^{-2N+1}\|f\|_{L^2(\Gamma)}\Big\} \\ &\text{for any } f \in C^\infty(\Gamma), z \in \Lambda_{a_0,1}, \mathrm{Im}\ z \ne 0,\end{aligned} \tag{6.2}$$

because $X(z) = (T^\pm(z))^{-1}B_{2N}(z) + (T^\pm(z))^{-1}(B_{2N'}(z) - B_{2N}(z)) + \gamma_\Gamma R^\pm(z)V_{2N'}(z)$ for any $N' > N$ and $B_{2N'}(\lambda + \sqrt{-1}\mu) - B_{2N}(\lambda + \sqrt{-1}\mu) \in \Psi^{-2N+1,-2N+1}_{0,0}(\Gamma : [1,\infty))$ uniformly in μ with $\lambda + \sqrt{-1}\mu \in \Lambda_{a_1,1}$. Putting z_j and f_j into (6.1) and using (2) in Theorem 6.1, we have

$$1 - C_{N'}|\mathrm{Re}\ z_j|^{-N'} \le C|\mathrm{Im}\ z_j|^{-1}\{C_{N'}|\mathrm{Re}\ z_j|^{-N'} + |\mathrm{Re}\ z_j|^{-2N+1}\},$$

which yields (3) since $z_j \in \Lambda_{a_0,a_1}$. This completes the proof of Theorem 6.1.

7. Null solutions of $P(z)$.

In this section, we show Proposition 6.2. Since the proof is not short, we divide it into three steps to obtain it.

Step 1. (a priori estimates). We show that there are constants $C > 0$ and $b_0 \ge a_1$ such that

$$\|f\|_{L^2(\Gamma)} \le C|\mathrm{Im}\ z|^{-1}|z|^{-2N+1}\|P(z)f\|_{L^2(\Gamma)}, \tag{7.1}$$

$$\|f\|_{L^2(\Gamma)} \le C|\mathrm{Im}\ z|^{-1}|z|^{-2N+1}\|(P(z))^*f\|_{L^2(\Gamma)} \tag{7.2}$$

$$\text{for any } f \in C^\infty(\Gamma), z \in \Lambda_{a_0,b_0}, |\mathrm{Im}\ z| \ge b_0,$$

where $(P(z))^*$ is the formal adjoint operator defined by the ordinary $L^2(\Gamma)$ inner product. Furthermore, there are constants C, $C' > 0$ such that

$$\|f\|_{H^{2N}(\Gamma)} \leq C\,\|P(z)f\|_{L^2(\Gamma)} + C'|\text{Re } z|^{2N}\,\|f\|_{L^2(\Gamma)} \quad \text{for any } f \in C^\infty(\Gamma), z \in \Lambda_{a_0,a_1}. \tag{7.3}$$

Proof of (7.1), (7.2), *and* (7.3). Choose a function $e \in C^\infty(T^*(\Gamma))$ satisfying $e(\zeta) = I$ near Σ, supp $e \subset W_0$. We introduce a cutoff operator E_λ ($\lambda > 1$) defined by

$$E_\lambda f(x) = \sum_{k=1}^{N_0} \phi_k(x)((s^k)^{-1})^*[Op_\lambda(e^k)(s^k)^*[\phi_k(\cdot)f(\cdot)]](x),$$

where $e^k(\sigma,\eta) = (\tilde{s}^k)^* e(\sigma,\eta)$. Here, we have used the same notations as in section 4.

By ellipticity of $P(z)$ on $T^*(\Gamma) \setminus \Sigma$, there is a pseudodifferential operator $Q_{\lambda,\mu} \in \Psi_{0,0}^{-2N,-2N}(\Gamma : [1,\infty))$ uniformly in μ with $\lambda + \sqrt{-1}\mu \in \Lambda_{a_0,a_1}$ such that

$$I - E_\lambda = Q_{\lambda,\mu} P(\lambda + \sqrt{-1}\mu) - R_{\lambda,\mu} \tag{7.4}$$

for some $R_{\lambda,\mu} \in \Psi_{0,0}^{-1,-1}(\Gamma : [1,\infty))$ uniformly in μ with $\lambda + \sqrt{-1}\mu \in \Lambda_{a_0,a_1}$. Thus, we have

$$\begin{aligned} \|(I - E_\lambda)f\|_{L^2(\Gamma)} \leq C|\lambda|^{-2N}\,&\|P(\lambda + \sqrt{-1}\mu)f\|_{L^2(\Gamma)} \\ &+ C'|\lambda|^{-1}\,\|f\|_{L^2(\Gamma)} \\ &\text{for any } f \in C^\infty(\Gamma), \lambda + \sqrt{-1}\mu \in \Lambda_{a_0,a_1}, \end{aligned} \tag{7.5}$$

with some constants $C > 0$ and $C' > 0$.

From the estimate (6.2) of $B_{2N}(z)$, we have

$$\begin{aligned} \|E_\lambda f\|_{L^2(\Gamma)} \leq C|\mu|^{-1}\,&\|E_\lambda B_{2N}(\lambda + \sqrt{-1}\mu)f\|_{L^2(\Gamma)} \\ &+ C'(|\lambda|^{-1} + |\mu|^{-1})\,\|f\|_{L^2(\Gamma)} \\ &\text{for any } f \in C^\infty(\Gamma), \lambda + \sqrt{-1}\mu \in \Lambda_{a_0,a_1}, \mu \neq 0. \end{aligned} \tag{7.6}$$

In fact, since $\sigma_p(X(\lambda + \sqrt{-1}\mu)E_\lambda) = (a - (\sqrt{-1}\mu/\lambda)H_{rad}(a)) \cdot e = e = \sigma_p(E_\lambda)$ by the property (3) of Proposition 4.2 and $a = I$ near W_0, and $[B_{2N}(\lambda + \sqrt{-1}\mu), E_\lambda] \in \Psi_{0,0}^{0,0}(\Gamma : [1,\infty))$, it follows that

$$\begin{aligned} \|E_\lambda f\|_{L^2(\Gamma)} &\leq \|X(\lambda + \sqrt{-1}\mu)E_\lambda f\|_{L^2(\Gamma)} + C'|\lambda|^{-1}\,\|f\|_{L^2(\Gamma)}, \\ \|B_{2N}(\lambda + \sqrt{-1}\mu)E_\lambda f\|_{L^2(\Gamma)} &\leq \|E_\lambda B_{2N}(\lambda + \sqrt{-1}\mu)f\|_{L^2(\Gamma)} \\ &\quad + C'\,\|f\|_{L^2(\Gamma)}. \end{aligned}$$

We also have $E_\lambda(I - X(\lambda + \sqrt{-1}\mu)) \in \Psi_{0,0}^{-\infty,-1}(\Gamma : [1,\infty))$, since supp $e \subset T^*(\Gamma)$ is compact and $\sigma_p(E_\lambda) - \sigma_p(E_\lambda X(\lambda + \sqrt{-1}\mu)) = 0$. From this fact, it follows that $E_\lambda(I - X(\lambda + \sqrt{-1}\mu))((\lambda + \sqrt{-1}\mu)^2 - \triangle_\Gamma)^N \in \Psi_{0,0}^{0,-1+2N}(\Gamma : [1,\infty))$, Thus, noting that $B_{2N}(z) = z^{-2N+1}(P(z) - \sqrt{-1}(I - X(z))(z^2 - \triangle_\Gamma)^N)$, by the estimate (7.6) we obtain

$$\begin{aligned} &\|E_\lambda f\|_{L^2(\Gamma)} \leq C|\mu|^{-1}|z|^{-2N+1}\,\|P(z)f\|_{L^2(\Gamma)} + C'(|\lambda|^{-1} + |\mu|^{-1})\,\|f\|_{L^2(\Gamma)} \\ &\quad \text{for any } f \in C^\infty(\Gamma), \lambda + \sqrt{-1}\mu \in \Lambda_{a_0,a_1}, \mu \neq 0. \end{aligned}$$

The estimate above and the estimate (7.5) yield (7.1).

To prove (7.2), we note that $(P(\lambda+\sqrt{-1}\mu))^*$ is also an elliptic operator on $T^*(\Gamma)\setminus\Sigma$ including every infinite point. Moreover, $(B_{2N}(z))^*$ has an estimate

$$\begin{aligned}\left\|X(\overline{z})f\right\|_{L^2(\Gamma)} \leq & C|\mathrm{Im}\ z|^{-1}\left\|(B_{2N}(z))^*f\right\|_{L^2(\Gamma)}\\ &+C(|\mathrm{Re}\ z|^{-1}+|\mathrm{Im}\ z|^{-1})\left\|f\right\|_{L^2(\Gamma)}\\ &\text{for any } f\in C^\infty(\Gamma), z\in\Lambda_{a_0,1}, \mathrm{Im}\ z\neq 0,\end{aligned}$$

by the property (2) in Proposition 4.2 and the estimate (6.2). Thus, the same argument to show (7.1) implies the estimate (7.2).

Since $Q_{\lambda,\mu}\in\Psi_{0,0}^{-2N,-2N}(\Gamma:[1,\infty))$, $E_\lambda\in\Psi_{0,0}^{-\infty,0}(\Gamma:[1,\infty))$, and $R_{\lambda,\mu}\in\Psi_{0,0}^{-1,-1}(\Gamma:[1,\infty))$ we have $\left\|Q_{\lambda,\mu}f\right\|_{H^s(\Gamma)}\leq C\left\|f\right\|_{L^2(\Gamma)}$, $\left\|E_\lambda f\right\|_{H^s(\Gamma)}\leq C|\lambda|^s\left\|f\right\|_{L^2(\Gamma)}$, and $\left\|R_{\lambda,\mu}f\right\|_{H^s(\Gamma)}\leq C\{\left\|f\right\|_{H^{s-1}(\Gamma)}+|\lambda|^{s-1}\left\|f\right\|_{L^2(\Gamma)}\}$ for any $f\in C^\infty(\Gamma)$, $\lambda+\sqrt{-1}\mu\in\Lambda_{a_0,a_1}$, $s=0,1,\ldots,2N$. Thus, from (7.4), it follows that

$$\begin{aligned}\left\|f\right\|_{H^s(\Gamma)}\leq & C\left\|P(\lambda+\sqrt{-1}\mu)f\right\|_{L^2(\Gamma)}+C(\left\|f\right\|_{H^{s-1}(\Gamma)}+|\lambda|^s\left\|f\right\|_{L^2(\Gamma)})\\ \text{for any } f\in & C^\infty(\Gamma), \lambda+\sqrt{-1}\mu\in\Lambda_{a_0,a_1}, s=0,1,\ldots,2N.\end{aligned}$$

Hence, by induction on s, we obtain (7.3).

Step 2. (realization of $P(z)$). As in section 6, we consider the operator $P(z): L^2(\Gamma)\to L^2(\Gamma)$, $D(P(z))=H^{2N}(\Gamma)$ for any $z\in\Lambda_{a_0,b_0}$. Here, we show that

(7.7) for any $z\in\Lambda_{a_0,b_0}$, $|\mathrm{Im}\ z|\geq b_0$, the operator $P(z)$ has the inverse $(P(z))^{-1}\in B(L^2(\Gamma),H^{2N}(\Gamma))$ such that for some fixed $C>0$, we have

$$\begin{aligned}\left\|(P(z))^{-1}f\right\|_{L^2(\Gamma)}\leq C|\mathrm{Im}\ z|^{-1}|z|^{-2N+1}\left\|f\right\|_{L^2(\Gamma)}\\ \text{for any } f\in L^2(\Gamma), z\in\Lambda_{a_0,b_0}, |\mathrm{Im}\ z|\geq b_0,\end{aligned}$$

(7.8) $(P(z))^{-1}$ is a $B(L^2(\Gamma),H^{2N}(\Gamma))$-valued finitely meromorphic function in $\mathbf{C}$,

(7.9) $$\begin{aligned}&\{\,z_0\in\mathbf{C}\,;\,z_0 \text{ is a pole of } (P(z))^{-1}\}\\ &\qquad=\{\,z_0\in\mathbf{C}\,;\,\mathrm{Ker}\ P(z_0)\neq\{0\}\,\}.\end{aligned}$$

Proof. To get (7.7), it suffices to check $P(z)$ is bijective. The estimate (7.1) says $P(z)$ is one-to-one. The estimates (7.1) and (7.3) imply the range $R(P(z))$ of $P(z)$ is closed in $L^2(\Gamma)$. If $0\neq f\in L^2(\Gamma)$ orthogonal to $R(P(z))$ exists, the function f have to satisfy $(P(z))^*f=0$ in $H^{-2N}(\Gamma)$ by duality, where $(P(z))^*$ is the realization of the formal adjoint operator of $P(z)$ by the inner product of $L^2(\Gamma)$. Thus, ellipticity of $(P(\lambda+\sqrt{-1}\mu))^*$ at every infinite point ensures $f\in C^\infty(\Gamma)$, which means $f=0$ because of the estimate (7.2). Hence, the operator $P(z)$ ($z\in\Lambda_{a_0,b_0}$, $|\mathrm{Im}\ z|\geq b_0$) is bijective.

To show (7.8), we decompose $P(z)$ as $P(z)=\sqrt{-1}(I-K(z))(I-\triangle_\Gamma)^N$, where $K(z)=K_1(z)+K_2(z)$, $K_1(z)=I-(I+(z^2-1)(I-\triangle_\Gamma)^{-1})^N=\sum_{j=1}^N\binom{N}{j}(z^2-1)^{N-j}(I-\triangle_\Gamma)^{-j}$, $K_2(z)=X(z)(I+(z^2-1)(I-\triangle_\Gamma)^{-1})^N+\sqrt{-1}z^{2N-1}B_{2N}(z)(I-\triangle_\Gamma)^{-N}$. From the form of $K(z)$ and compactness of Γ, it follows that $K(z)$ is a $B(L^2(\Gamma))$-valued entire function and is a compact operator for any $z\in\mathbf{C}$. By (7.7), $-\sqrt{-1}P(z)(I-\triangle_\Gamma)^{-N}\in B(L^2(\Gamma))$ has the inverse for $z\in\Lambda_{a_0,b_0}$, $|\mathrm{Im}\ z|\geq b_0$.

It means that there exists the inverse $(I - K(z))^{-1} \in B(L^2(\Gamma))$ for $z \in \Lambda_{a_0,b_0}$, $|\mathrm{Im}\ z| \geq b_0$. The analytic Fredholm theorem implies $(I - K(z))^{-1}$ is a $B(L^2(\Gamma))$-valued finitely meromorphic function in $\mathbf{C}$ whose poles are discrete. This completes the proof of (7.8). From (7.8), the property (7.9) is obvious.

Step 3. (deduction of inconsistency). It suffices to prove that for any fixed $b \geq b_0$, there exists $z_0 \in \Lambda_{a_0,b}$ such that Ker $P(z_0) \neq \{0\}$. In this step, we show it by deriving a contradiction.

Assume that there is a constant $b \geq b_0$ satisfying Ker $P(z) = \{0\}$ for any $z \in \Lambda_{a_0,b}$ as the hypothesis in the present contradiction argument. From (7.8), (7.9), the hypothesis implies the operator $(P(z))^{-1}$ is holomorphic in $\Lambda_{a_0,b}$. Then, by the argument proving Lemma 3 of Stefanov and Vodev [21], we also have the estimate

$$\left\|(P(z))^{-1}\right\| \leq Ce^{C|z|^{n+1}} \quad \text{on } \Lambda_{a_0/2,b'}, \tag{7.10}$$

with some fixed constants $C > 0$ and $b' \geq b$. Since the operator $z^{2N-1}(\log z)(P(z))^{-1}$ is uniformly bounded in the boundary of $\Lambda_{a_0/2,b'}$ as a $B(L^2(\Gamma))$-valued function because of the estimate (7.1), we obtain

$$\left\|(P(z))^{-1}\right\|_{B(L^2(\Gamma))} \leq C|\log z|^{-1}|z|^{-2N+1} \quad \text{on } \Lambda_{a_0/2,b'}.$$

In fact, we can use the Phragmén and Lindelöf theorem to the operator $z^{2N-1}(\log z)$ $(P(z))^{-1}$ by the a priori estimate (7.10). Thus, from the hypothesis of the contradiction argument, it follows that there exist constants $C > 0$ and $b' \geq b$ such that

$$\begin{aligned}\left\|f\right\|_{L^2(\Gamma)} &\leq C|\log z|^{-1}|z|^{-2N+1}\left\|P(z)f\right\|_{L^2(\Gamma)} \\ &\text{for any } f \in C^\infty(\Gamma), z \geq b'.\end{aligned} \tag{7.11}$$

Next, we show the estimate (7.11) is incompatible with Proposition 5.2 deduced from the assumption (ERW). For the local coordinate $s : \widetilde{U} \to U$ in Proposition 5.2, we choose functions ϕ, $\psi_j \in C_0^\infty(\widetilde{U})$, $(j = 1, 2)$ satisfying $\phi(\sigma) = 1$ near $\sigma = \sigma^0$, $\psi_1(\sigma) = 1$ near supp ϕ, $\psi_2(\sigma) = 1$ near supp ψ_1, we set $\Psi_j(x) = (s^{-1})^*\psi_j(x)$ $(j = 1, 2)$. From supp $(I - \Psi_2) \cap$ supp $\Psi_1 = \phi$, it follows $\left\|(I - \Psi_2)P(z)\Psi_1\right\|_{B(L^2(\Gamma))} \leq C_{N'}|z|^{-N'}$ for any $z \geq 1$. Thus, by the estimate (7.11), we have

$$\begin{aligned}\left\|\phi u\right\|_{L^2(\mathbf{R}_\sigma^{n-1})} &\leq C|z|^{-2N+1}|\log z|^{-1}\left\|\widetilde{P}(z)\phi u\right\|_{L^2(\mathbf{R}_\sigma^{n-1})} \\ &\qquad + C_{N'}|z|^{-N'}\left\|\phi u\right\|_{L^2(\mathbf{R}_\sigma^{n-1})} \\ &\text{for any } u \in C^\infty(\mathbf{R}_\sigma^{n-1}), z \geq b',\end{aligned} \tag{7.12}$$

where $\widetilde{P}(z)v(\sigma) = (s^*\Psi_2 P(z)\Psi_1(s^{-1})^*v)(\sigma)$ $(v \in C_0^\infty(\mathbf{R}_\sigma^{n-1}))$. Note that $\widetilde{P}(z) \in \Psi_{0,0}^{2N,2N}(\mathbf{R}^{n-1} : [1, \infty))$ and $\sigma_p(\widetilde{P}(z))(\sigma, \eta) = z^{2N}\psi_1(\sigma)\{(\tilde{s}^*a)(\sigma, \eta)\ (\tilde{s}^*l_0)(\sigma, \eta) + \sqrt{-1}$ $(1 - (\tilde{s}^*a)(\sigma, \eta))(1 + |\eta|_\Gamma^2)^N\}$.

Since $\widetilde{W}$ in Proposition 5.2 meets $\tilde{s}^{-1}(W_0)$, for $\tilde{s}((\sigma^0, \eta^0)) \in \Sigma \subset W_0$, we can choose a function $a_0(\sigma, \eta) \in C_0^\infty(\widetilde{U} \times \mathbf{R}^{n-1})$ satisfying $a_0(\sigma, \eta) = 1$ near (σ^0, η^0), supp $a_0 \subset \tilde{s}^{-1}(W_0) \cap \widetilde{W} \cap ($(supp $\psi_2) \times \mathbf{R}^{n-1})$. We set $\tilde{q}(\sigma, \eta) = a_0(\sigma, \eta)q(\sigma, \eta)$, where $q(\sigma, \eta)$ is the function in Proposition 5.2. Note that $\widetilde{P}(z)\phi Op_z(\tilde{q}) \in \Psi_{0,0}^{2N,2N}(\mathbf{R}^{n-1} :$ $[1, \infty))$ and $\sigma_p(\widetilde{P}(z)\phi Op_z(\tilde{q}))(\sigma, \eta) = z^{2N}\phi(\sigma)a_0(\sigma, \eta)(\tilde{s}^*l_0)(\sigma, \eta)q(\sigma, \eta)$.

The property $\lambda(\sigma^0, \eta^0) = 1$, $(\partial_\eta\lambda)(\sigma^0, \eta^0) \neq 0$ for the function λ in Proposition 5.2 ensures existence of a C^∞-function $S(\sigma)$ in an open neighborhood $\widetilde{V} \subset \widetilde{U}$ of σ^0

satisfying $\lambda(\sigma, \nabla_\sigma S(\sigma)) = 1$ in $\widetilde{V}$, $(\nabla_\sigma S)(\sigma^0) = \eta^0$. For the function $S(\sigma)$ and any function $\varphi \in C_0^\infty(\widetilde{V})$, we consider a function $\exp(\sqrt{-1}zS(\sigma))\varphi(\sigma)$ $(z \geq 1)$. From the expanding theorem of pseudodifferential operator and $\sigma_p(\widetilde{P}(z)\phi Op_z(\tilde{q}))(\sigma, \nabla_\sigma S(\sigma))$ $= z^{2N}\phi(\sigma)a_0(\sigma,\eta)\ (\tilde{s}^* l_0)(\sigma,\eta)q(\sigma,\eta)|_{\eta=\nabla_\sigma S(\sigma)} = 0$ by Proposition 5.2, it follows that

$$\left\|\widetilde{P}(z)\phi Op_z(\tilde{q})(e^{\sqrt{-1}zS(\cdot)}\varphi(\cdot))\right\|_{L^2(\mathbf{R}_\sigma^{n-1})} \leq C|z|^{2N-1},$$
$$\left\|Op_z(\tilde{q})(e^{\sqrt{-1}zS(\cdot)}\varphi(\cdot)) - e^{\sqrt{-1}zS(\cdot)}\tilde{q}(\cdot, \nabla_\sigma S(\cdot))\varphi(\cdot)\right\|_{L^2(\mathbf{R}_\sigma^{n-1})} \leq C|z|^{-1}$$

for any $z \geq 1$ with a fixed constant $C > 0$. Hence, putting the function $u(\sigma) = Op_z(\tilde{q})(e^{\sqrt{-1}zS(\cdot)}\varphi(\cdot))(\sigma)$ into the estimate (7.12), we obtain

$$\left\|\phi(\cdot)\tilde{q}(\cdot, \nabla_\sigma S(\cdot))\varphi(\cdot)\right\|_{L^2(\mathbf{R}_\sigma^{n-1})} \leq C|\log z|^{-1} \qquad \text{for any } z \geq b',$$

which yields $q(\sigma^0, \eta^0) = 0$ by taking the limit as $z \to \infty$. This conclusion, however, make an inconsistency because $q(\sigma^0, \eta^0) \neq 0$ as in Proposition 5.2. This completes the proof of Proposition 6.2.

Appendix A. Meromorphic continuation of the resolvent. Here, we give a proof of properties of $R^\pm(z)$ used throughout in the present paper. We use notations introduced in the Introduction.

Theorem A.1. *Assume that* (A.1) *and* (A.2) *are satisfied and that every coefficient* $C^1_{ijkl}(x)$ *belongs to the space* $C_0^\infty(\mathbf{R}^n)$. *Then, the resolvent* $R^\pm(z)$ *can be continued as a* $B(L_a^2(\Omega), H^2(\Omega_a))$*-valued finitely meromorphic function in* $\widetilde{\mathbf{C}}_\pm$, *and the set consisting of poles of* $R^\pm(z)$ *is a discrete set.*

To prove Theorem A.1, we need the properties of the outgoing (resp., incoming) resolvent $R_0^+(z)$ (resp., $R_0^-(z)$) for the free space problem, that is,

$$\begin{cases} (A^0(\partial_x) + z^2)v(x;z) = f(x) & \text{in} \quad \mathbf{R}^n, \\ v(x;z) \text{ is outgoing (resp., incoming).} \end{cases}$$

Note that $R_0^\pm(z)$ is a $B(L^2(\mathbf{R}^n), H^2(\mathbf{R}^n))$-valued holomorphic function in $\pm\text{Im } z < 0$.

Theorem A.2. *Assume that* (A.1) *and* (A.2) *hold. Then for any* $a > 0$, $R_0^\pm(z)$ *is continued analytically in* $\widetilde{\mathbf{C}}_\pm$ *as* $B(L_a^2(\Omega), H^2(\Omega_a))$*-valued function. Furthermore, there exist constants* $C_a > 0$, $T_a > 0$ *such that*

$$\left\|R_0^\pm(z)f\right\|_{H^{2-j}(B_a)} \leq C_a|z|^{1-j}e^{T_a|\text{Im } z|}\left\|f\right\|_{L^2(\mathbf{R}^n)}$$
$$\textit{for any } f \in L_a^2(\mathbf{R}^n), z \in \widetilde{\mathbf{C}}_\pm \textit{ and } j = 0, 1, 2.$$

Proof of Theorem A.2. Consider the Cauchy problem for the operator $\partial_t^2 - A^0(\partial_x)$,

$$\text{(a.1)} \qquad \begin{cases} (\partial_t^2 - A^0(\partial_x))u(t,x) = 0 & \text{in} \quad \mathbf{R} \times \mathbf{R}^n, \\ u(0,x) = 0, \quad \partial_t u(0,x) = f(x) & \text{on} \quad \mathbf{R}^n. \end{cases}$$

Since $R_0^\pm(z)f(x) = -\int_0^\infty \exp(\mp\sqrt{-1}zt)u(t,x)\,dt$ for $\pm\text{Im } z < 0$, to show Theorem A.2, it is important to know the properties of the solution of (a.1).

Set $C_{min} = \inf_{\omega \in S^{n-1}} \min\{(A(\omega)x, x)_{\mathbf{C}^n}\ ;\ x \in \mathbf{C}^n, |x| = 1\} > 0$, where $(\cdot,\cdot)_{\mathbf{C}^n}$ means the standard inner product of $\mathbf{C}^n$.

Theorem A.3. *Assume that* (A.1) *and* (A.2) *are satisfied. Choose fixed* $a > 0$. *Then the solution of the Cauchy problem* (a.1) *satisfies*

(i) *in odd n case,* $u(t,x) = 0$ *in* $t > C_{min}^{-1}(a+|x|)$ *if* $f \in L_a^2(\mathbf{R}^n)$,

(ii) *in even n case,* $u(t,x)$ *is* C^∞ *in* $t > C_{min}^{-1}(a+|x|)$ *and can be continued analytically in t to the region* $\operatorname{Re} t > C_{min}^{-1}(a+|x|)$ *if* $f \in L_a^2(\mathbf{R}^n)$. *Moreover, for any* $\epsilon > 0$, *we have an estimate*

$$|(\partial_t^l \partial_x^\alpha u)(t,x)| \le C_{l,\alpha,\epsilon}(1+|t|)^{-(n-1)-l-|\alpha|} \left\|f\right\|_{L^2(\mathbf{R}^n)}$$
$$\text{for all } f \in L_a^2(\mathbf{R}^n), \operatorname{Re} t > C_{min}^{-1}(a+|x|)+\epsilon,$$

with a constant $C_{l,\alpha,\epsilon} > 0$ *depending only on l, α, and* $\epsilon > 0$.

According to Vainberg's argument (cf. [24]), we can also prove Theorem A.2 from Theorem A.3. Thus, to obtain Theorem A.2, it suffices to show Theorem A.3.

Proof of Theorem A.3. By the argument in Wilcox [27], there is a positive integer d $(1 \le d \le n)$ such that all eigenvalues of $A^0(\xi) = \sum_{i,l=1}^n C_{il}^0 \xi_i \xi_l$ can be enumerated as $0 < \lambda_1(\xi) \le \lambda_2(\xi) \le \cdots \le \lambda_d(\xi)$. Each $\lambda_j(\xi)$ is a continuous even function, homogeneous of order 2 and $\lambda_j(\xi) \ge C_{min}|\xi|^2$ for any $\xi \in \mathbf{R}^n$ $(j = 1, \ldots, d)$.

Wilcox [25] also shows that there exists a closed conic set $\mathcal{O} \subset \mathbf{R}^n$ which is measure zero in $\mathbf{R}^n$, that is, $S^{n-1} \cap \mathcal{O}$ is measure zero in S^{n-1}, such that

$$\lambda_1(\xi) < \lambda_2(\xi) < \cdots < \lambda_d(\xi) \quad \text{for any } \xi \in \mathbf{R}^n \setminus \mathcal{O}, \tag{a.2}$$

and each $\lambda_j(\xi)$ is real analytic in $\mathbf{R}^n \setminus \mathcal{O}$.

From (a.2), we can define the eigenprojector $P_j(\xi)$ corresponding to the eigenvalues $\lambda_j(\xi)$ as $P_j(\xi) = (2\pi\sqrt{-1})^{-1} \int_{|\zeta - \lambda_j(\xi)| = \gamma_j(\xi)} (\zeta - A^0(\xi))^{-1}\, d\zeta$, where $\gamma_j(\xi) = 3^{-1} \inf\{|\lambda_i(\xi) - \lambda_j(\xi)|\ ;\ i \ne j, i, j = 1, \ldots, d\} > 0$. By definition, every $P_j(\xi)$ is real analytic in $\mathbf{R}^n \setminus \mathcal{O}$, homogenous of order 0, and even function. For any $\xi \in \mathbf{R}^n \setminus \mathcal{O}$, they satisfy $A^0(\xi) = \sum_{i=1}^d \lambda_i(\xi) P_i(\xi)$ and

$$P_i(\xi)P_j(\xi) = \delta_{ij} P_j(\xi),\ P_j(\xi) = (P_j(\xi))^*,\ \sum_{j=1}^d P_j(\xi) = I. \tag{a.3}$$

Since the property (a.3) implies $\left\|P_j(\xi)\right\|_{B(\mathbf{C}^n)} = 1$ for any $\xi \in \mathbf{R}^n \setminus \mathcal{O}$, $j = 1, \ldots, d$, each $P_j(\xi)$ is bounded measurable function in $\mathbf{R}^n$. Hence, for any $f \in C_0^\infty(\mathbf{R}^n)$, we can represent the solution $u(t,x)$ in the form

$$u(t,x) = (2\pi)^{-n} \int_{\mathbf{R}^n} e^{\sqrt{-1}\xi \cdot x} \sum_{j=1}^d \frac{\sin(t\mu_j(\xi))}{\mu_j(\xi)} P_j(\xi) \hat{f}(\xi)\, d\xi, \tag{a.4}$$

where $\mu_j(\xi) = \sqrt{\lambda_j(\xi)}$, $\hat{f}(\xi) = \int_{R^n} \exp(-\sqrt{-1}\xi \cdot x) f(x)\, dx$. In fact, the function defined by the integral in (a.4) is well defined as a C^∞-function and satisfies (a.1). Changing the integral in (a.4) to the poler coordinate, we obtain

$$\begin{aligned} u(t,x) = 2^{-1}(2\pi)^{1-n}\sqrt{-1} \sum_{j=1}^d \int_{S^{n-1}} \mu_j(\omega)^{-1} P_j(\omega),& \\ \mathcal{F}^{-1}[|p|^{n-2}\kappa(p)(\mathcal{F}\mathcal{R}f)](\omega \cdot x - t\mu_j(\omega), \omega)\, d\omega& \end{aligned} \tag{a.5}$$

because of the property $\hat{f}(p\omega) = (\mathcal{F}\mathcal{R}f)(p,\omega)$ (cf. section 2 in [17]), where $\mathcal{R}$ is the Radon transform defined by

$$(\mathcal{R}g)(s,\omega) = \int_{x \cdot \omega = s} g(x)\, dS_x \ \text{ for } \ (s,\omega) \in \mathbf{R} \times S^{n-1}$$

and $(\mathcal{F}k)(p) = \int_{\mathbf{R}} \exp(-\sqrt{-1}ps)k(s)\,ds$ is the Fourier transform. We use the representation (a.5) to show Theorem A.3.

First, we consider the odd dimensional case. If this is the case, we have $\mathcal{F}^{-1}[|\cdot|^{n-2}\kappa(\cdot)(\mathcal{F}\mathcal{R}f)](s,\omega) = (-\sqrt{-1}\partial_s)^{n-2}\mathcal{R}f(s,\omega)$. Since $\mathcal{R}f(s,\omega) = 0$ in $|s| \geq a$ for any $f \in C_0^\infty(B_a)$, the representation (a.5) implies $u(t,x) = 0$ for any (t,x) satisfying $|\omega \cdot x - t\mu_j(\omega)| \geq a$ for all $\omega \in S^{n-1}$. Hence, we have the statement (i) in Theorem A.3 with assumption $f \in C_0^\infty(B_a)$. In fact,

$$|\omega \cdot x - t\mu_j(\omega)| \geq C_{\min}t - |x| > a \text{ if } t > C_{\min}^{-1}(|x| + a). \tag{a.6}$$

Noting that problem (a.1) is well posed in $C^1(\mathbf{R} : L^2(\mathbf{R}^n)) \cap C(\mathbf{R} : H^1(\mathbf{R}^n))$, we obtain Theorem A.3 in the odd n case.

Second, we consider the even dimensional case. Note that

$$\mathcal{F}^{-1}[|\cdot|^{n-2}\kappa(\cdot)(\mathcal{F}\mathcal{R}f)](p,\omega) = (\sqrt{-1})^{3-n}\partial_p^{n-2}(\mathcal{H}\mathcal{R}f)(p,\omega), \tag{a.7}$$

where $(\mathcal{H}k)(p) = \pi^{-1}\lim_{\epsilon\to 0}\int_{|p'|\geq\epsilon} k(p-p')/p'\,dp'$ is the Hilbert transform. From (a.6) and (a.7), it follows that

$$u(t,x) = c_n \sum_{j=1}^{d} \int_{S^{n-1}} \mu_j(\omega)^{-1}P_j(\omega) \int_{|p'|\leq a} \frac{(\mathcal{R}f)(p',\omega)}{(\omega\cdot x - \mu_j(\omega)t - p')^{n-1}}\,dp'\,d\omega$$
$$\text{for any } t > C_{\min}^{-1}(a + |x|),\ f \in C_0^\infty(B_a),$$

with some constant $c_n \in \mathbf{C}$. The representation implies the statement (ii) if $f \in C_0^\infty(B_a)$, because $\left\|\mathcal{R}f(\cdot,\omega)\right\|_{L^2(\mathbf{R}^n)} \leq Ca^{(n-1)/2}\left\|f\right\|_{L^2(\mathbf{R}^n)}$ for any $\omega \in S^{n-1}$, $f \in C_0^\infty(B_a)$ with some constant $C > 0$. Hence, by a density argument, we obtain (ii) in Theorem A.3. This completes the proof of Theorem A.3. Hence, we also have Theorem A.2.

Proof of Theorem A.1. We can write down $R_0^\pm(z)$ in $\pm\text{Im } z < 0$ as

$$R_0^\pm(z)f(x) = (2\pi)^{-n}\int_{\mathbf{R}^n} e^{\sqrt{-1}x\cdot\xi}\sum_{j=1}^{d}\frac{P_j(\xi)\hat{f}(\xi)}{z^2 - \lambda_j(\xi)}\,d\xi \quad (f \in L^2(\mathbf{R}^n))$$

because the properties of $\lambda_j(\xi)$ and $P_j(\xi)$ imply well-definedness of the function defined by the integral above and it gives us an L^2-solution of $(A^0(\partial_x) + z^2)v(x;z) = f(x)$. From the expression and the argument in section 2 of Iwashita and Shibata [9], it follows that for the free space resolvent $R_0^\pm(z)$, $R_0^\pm(0)$ is well defined and $R_0^\pm(z)$ is continuous on $\{0\} \cup \widetilde{\mathbf{C}}_\pm$ since the operators defined by multiplier $|\xi|^2P_j(\xi)/\lambda_j(\xi)$ are $L^2(\mathbf{R}^n)$ bounded. Hence, using analytic Fredholm theory like as in section 2 of [9], we can obtain Theorem A.1.

Acknowledgments. A part of the work was done when the first author was staying at UCLA in 1995. The first author would like to thank Professor James Ralston for his kindness and hospitality and to thank UCLA for the pleasant conditions. The first author would also like to show appreciation to his colleagues in the Faculty of Education of Ibaraki University for allowing him to be absent.

REFERENCES

[1] J.D. Achenbach, *Wave Propagation in Elastic Solids*, North–Holland, Amsterdam, 1973.

[2] D.M. Barnett and J. Lothe, *Free surface (Rayleigh) waves in anisotropic elastic half spaces: The surface impedance method*, Proc. Roy. Soc. London Ser. A, 402 (1985), pp. 135–152.

[3] P. Chadwick and G.B. Smith, *Foundations of the theory of surface waves in anisotropic elastic materials*, in Advances in Applied Mechanics 17, Academic Press, New York, 1977, pp. 303–375.

[4] C. Gérard, *Asymptotique des poles de la matrice de scattering pour deux obstacles strictement convexes*, Bull. Soc. Math. France, 116 (1988).

[5] M. Ikawa, *On the poles of the scattering matrix for two strictly convex obstacles*, J. Math. Kyoto Univ., 27 (1983), pp. 127–194.

[6] M. Ikehata and G. Nakamura, *Decaying and nondecaying properties of the local energy of an elastic wave outside an obstacle*, Japan J. Appl. Math., 6 (1989), pp. 83–95.

[7] H. Ito, *Extended Korn's inequalities and the associated best possible constants*, J. Elasticity, 24 (1990), pp. 43–78.

[8] H. Iwashita, *A remark on the analyticity of spectral functions for some exterior boundary value problem*, Sci. Rep. Niigata Univ. Ser. A, 24 (1988), pp. 25–31.

[9] H. Iwashita and Y. Shibata, *On the analyticity of spectral functions for exterior boundary value problems*, Glas. Mat. Ser. III, 23 (1988), pp. 291–313.

[10] M. Kawashita, *On the local energy decay property for the elastic wave equation with the Neumann boundary condition*, Duke Math. J., 67 (1992), pp. 333–351.

[11] M. Kawashita, *On the decay rate of local energy for the elastic wave equation*, Osaka J. Math., 30 (1993), pp. 813–837.

[12] M. Kawashita, *On a region free from the poles of the resolvent and decay rate of the local energy for the elastic wave equation*, Indiana Univ. Math. J., 43 (1994), pp. 1013–1043.

[13] H. Kumano-go, *Pseudo-Differential Operators*, MIT Press, Cambridge, MA, London, 1982.

[14] C.S. Morawetz, *Exponential decay of solutions of the wave equation*, Comm. Pure Appl. Math., 19 (1966), pp. 439–444.

[15] G. Nakamura, *Existence and propagation of Rayleigh waves and pulses*, in Modern Theory of Anisotropic Elasticity and Applications, SIAM, Philadelphia, 1991, pp. 215–231.

[16] J. Ralston, *Solutions of the wave equation with localized energy*, Comm. Pure Appl. Math., 22 (1969), pp. 807–823.

[17] Y. Shibata and H. Soga, *Scattering theory for the elastic wave equation*, Publ. Res. Inst. Math. Sci., 25 (1989), pp. 861–887.

[18] J. Sjöstrand and G. Vodev, *Asymptotics of the number of Rayleigh resonances*, Math. Ann., 309 (1997), pp. 287–306.

[19] P. Stefanov and G. Vodev, *Distribution of the resonances for the Neumann problem in linear elasticity in the exterior of a ball*, Ann. Inst. H. Poincaré Phys. Théor., 60 (1994), pp. 303–321.

[20] P. Stefanov and G. Vodev, *Distribution of the resonances for the Neumann problem in linear elasticity in the exterior of a strictly convex body*, Duke Math. J., 78 (1995), pp. 677–714.

[21] P. Stefanov and G. Vodev, *Neumann resonances in linear elasticity for an arbitrary body*, Comm. Math. Phys., 176 (1996), pp. 645–659.

[22] A.N. Stroh, *Steady state problems in anisotropic elasticity*, J. Math. Phys., 41 (1962), pp. 77–103.

[23] M. Taylor, *Rayleigh waves in linear elasticity as a propagation of singularities phenomenon*, in Partial Differential Equations and Geometry, Marcel Dekker, New York, 1979, pp. 273–291.

[24] B.R. Vainberg, *On the short wave asymptoticbehavior of solutions of stationary problems and the asymptoticbehavior as $t \to \infty$ of non-stationary problems*, Russian Math. Surveys, 30 (1975), pp. 1–58.

[25] G. Vodev, *Existence of Rayleigh resonances exponentially close to the real axis*, Ann. Inst. H. Poincaré Phys. Théor., 67 (1997), pp. 41–57.

[26] H.F. Walker, *Some remarks on the local energy decay of solutions of the initial-boundary value problem for the wave equation in unbounded domains*, J. Differential Equation, 23 (1977), pp. 459–471.

[27] C. Wilcox, *Asymptotic wave functions and energy distributions in strongly propagative anisotropic media*, J. Math. Pures Appl., 57 (1978), pp. 275–321.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 726–739

ON THE SPLITTING TRICK AND WAVELET FRAME PACKETS[*]

DI-RONG CHEN[†]

Abstract. This paper presents a detailed analysis of the "splitting trick" which splits, for example, the half-shifts of a function into the shifts of two functions. When a Riesz basis of a shift-invariant subspace is split, the optimal bounds of the resulting Riesz basis are obtained. Most importantly, by the splitting trick we built wavelet frame packets as orthogonal wavelet packets constructed by Coifman and Meyer. Their algorithms for finding best basis for a function also apply to our setting.

Key words. frame, wavelet, wavelet frame, Riesz basis, shift-invariant space

AMS subject classification. 42C15

PII. S0036141097323333

1. Introduction.

The aim of this paper is to construct wavelet frame packets, in which there are many frames. It is a generalization of wavelet packets. Meanwhile the efficient algorithms for finding best basis in wavelet packets apply to wavelet frame packets.

Let $\mathcal{H}$ be a separable Hilbert space. A sequence $\{f_k\}_{k=1}^{\infty} \subseteq \mathcal{H}$ is called a frame of $\mathcal{H}$ if there are constants A and B, $0 < B \leq A < \infty$ such that

$$B\|f\|^2 \leq \sum_{k=1}^{\infty} |\langle f, f_k\rangle|^2 \leq A\|f\|^2 \quad \forall f \in \mathcal{H}, \tag{1.1}$$

where $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ are the norm and inner product of $\mathcal{H}$, respectively. The frame bounds are the smallest A and largest B that can be used in (1.1). If we can choose $A = B$ in (1.1), then $\{f_k\}_{k=1}^{\infty}$ is a tight frame with bound A. In particular, when $A = B = 1$, we have the reconstruction formula

$$f = \sum_{k=1}^{\infty} \langle f, f_k\rangle f_k, \tag{1.2}$$

which looks exactly as in the orthonormal basis case.

A frame of $L_2 := L_2(\mathbb{R}^s)$ of form $\{2^{k/2}\psi(2^k\cdot{-\alpha}) \mid k \in \mathbb{Z}, \alpha \in \mathbb{Z}^s, \psi \in \Psi\}$ is called a wavelet frame, where Ψ is a finite set of $L_2(\mathbb{R}^s)$. In [11], some explicit constructions of wavelet frames and some conditions ensuring the existence of such frames were given.

If a wavelet frame $\{2^{k/2}\psi(2^k\cdot -\alpha) \mid k \in \mathbb{Z}, \alpha \in \mathbb{Z}^s, \psi \in \Psi\}$ is an orthonormal basis of L_2, Ψ is called a (orthonormal) wavelet set. Usually, a wavelet set is derived from a given multiresolution analysis of L_2. The construction of wavelet sets has been

[*] Received by the editors June 24, 1997; accepted for publication (in revised form) February 25, 1999; published electronically March 21, 2000. This work was supported in part by National Natural Science Foundation of China under grants 195005 and 198003. The first version of the manuscript was written when the author was a postdoctoral fellow of Institute of Mathematics, Chinese Academy of Sciences.

http://www.siam.org/journals/sima/31-4/32333.html

[†] Department of Applied Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100083, People's Republic of China (drchen@263.net).

discussed in a great number of papers. The problem has been settled more successfully in case $s = 1$ than $s > 1$. One of the difficulties in the latter case is that there is not a common and useful method to solve the problem of matrix extension; see, e.g., [1], [4], [14], [15], [18], and [19].

For a given multiresolution analysis and the corresponding orthonormal wavelet basis of $L_2(\mathbb{R}^s)$, wavelet (basis) packets were constructed by Coifman and Meyer. This construction is an important generalization of that of wavelet sets. There are many orthonormal bases in the wavelet packets. Efficient algorithms for finding the best possible basis do exist (in the context of signal analysis application); however, as pointed out in [11] and [12, pp. 97–99], for certain wavelet applications in signal analysis, frames are more suitable than orthonormal bases, due to the redundancy in frames. It is therefore worthwhile to generalize the construction of wavelet packets to the frame case.

Indeed, orthogonality cannot always be considered to be a crucial property of the wavelet basis [12, Chapter 8], and there has been an extensive study of other wavelet systems, such as prewavelets and biorthogonal wavelets ([1], [5], [7], [9], [14], [17], etc.) as well as nonorthogonal wavelet packets ([6] and [8]). These sequences are Riesz bases of L_2.

A sequence $\{f_k\}_{k=1}^{\infty} \subseteq \mathcal{H}$ is a Riesz basis for $\mathcal{H}$ if there are some constants A and B, $0 < B \leq A < \infty$, such that any $f \in \mathcal{H}$ can be represented as a series $f = \sum_{k=1}^{\infty} c_k f_k$ converging in $\mathcal{H}$ with

$$B\|f\|^2 \leq \sum_{k=1}^{\infty} |c_k|^2 \leq A\|f\|^2. \tag{1.3}$$

The Riesz basis bounds are the smallest A and largest B that can be used in (1.3).

We recall that a Riesz basis with bounds A and B is a frame with bounds A and B. Conversely, a frame $\{f_k\}_{k=1}^{\infty}$ is a Riesz basis provided $f_1, f_2, \ldots$ are linearly independent in the following sense: any f_{k_0} does not belong to the closure of all finitely linear combinations of elements in $\{f_k\}_{k \neq k_0}$.

The main tool used in wavelet packets is the splitting trick (Daubechies' terminology; cf. [12]), which breaks the half integer shifts, for example, of a function into the integer shifts of two functions. One of the basic results in [6] and [8] says that, even in the nonorthogonal case, as long as finitely many steps of splitting are applied, the resulting sequence is still a Riesz basis. On the other hand, the existing estimates for corresponding bounds are known to be suboptimal (cf. [8]). Moreover, the methods of [6] and [8] do not apply to the frame case. In any event, the condition number of the resulting Riesz basis may be large: a result in [8] shows that infinitely many splitting steps may lead to a system that is not a Riesz basis any more.

The paper is organized as follows. We review some basic concepts and results about shift-invariant spaces and prove a characterization of a frame in shift-invariant spaces in section 2. In section 3 we analyze in detail the splitting trick in shift-invariant spaces and obtain the sharp inequalities that correspond to this process. With these inequalities it can be proved easily that, if we apply at most L splittings to a frame (Riesz basis, respectively) of a shift-invariant space, the resulting sequence is still a frame (Riesz basis, respectively). In section 4 we consider the splitting trick on frames (Riesz bases, respectively) of L_2. We shall prove that as long as finitely many splitting steps are applied, the resulting sequence is a frame (Riesz basis, respectively) of L_2. In particular, if the matrix associated with the splitting is unitary, then the resulting sequence is a frame (Riesz basis, respectively) with bounds $A' \leq A$ and $B' \geq B$, where A and B are the bounds of the original sequence. In this case the splitting can be applied at infinitely many times.

We wish to compare our work with [6] and [8]. First, we work in a more general setting. In fact we don't need refinability of the underlying functions as well as a multiresolution. In our deduction only the properties of shift-invariant spaces are employed. Those properties were developed by de Boor, DeVore, and Ron and used extensively in multiresolution analysis, wavelets, and approximation theory. Second, even in the Riesz basis case, our estimation for the bounds of the resulting Riesz bases is sharper than that in either [6] or [8].

In the rest of the paper we assume that $\mathcal{H}$ is either L_2 or a subspace of $L_2(\mathbb{R}^s)$. Moreover, for simplicity of notations, we present only results for $L_2(\mathbb{R})$. However, all conclusions have their counterparts in $L_2(\mathbb{R}^s), s > 1$.

2. Shift-invariant spaces. For later use we review briefly some basic concepts and results for shift-invariant spaces. Then we prove a characterization for the integer shifts of a finite set to be a frame.

A linear space S of functions from $\mathbb{R}$ to $\mathbb{C}$ is called 2^{-k} shift-invariant if it is invariant under 2^{-k} shifts; that is,

$$f \in S \ \Rightarrow f(\cdot - 2^{-k}\alpha) \in S \quad \forall \alpha \in \mathbb{Z}.$$

For a finite set $\Phi \subseteq L_2 := L_2(\mathbb{R})$, we denote by $\mathcal{S}_0^k(\Phi)$ the (finite) linear span of the 2^{-k} shifts of the functions in Φ. Thus a function $f \in \mathcal{S}_0^k(\Phi)$ has the form

$$f = \sum_{\varphi \in \Phi} \sum_{\alpha \in \mathbb{Z}} c_{\varphi\alpha} \varphi(\cdot - 2^{-k}\alpha),$$

where the sequence of coefficients $\{c_{\varphi\alpha}\}_{\varphi \in \Phi, \alpha \in \mathbb{Z}}$ is finitely supported. Obviously $\mathcal{S}_0^k(\Phi)$ is the smallest 2^{-k} shift-invariant space containing Φ.

If Φ is a subset of L_2, we write $\mathcal{S}^k(\Phi)$ for the closure of $\mathcal{S}_0^k(\Phi)$ in L_2. We use the notation $\mathcal{S}^k(\varphi)$ instead of $\mathcal{S}^k(\{\varphi\})$ when Φ consists of a single function φ.

The following characterization of $\mathcal{S}^k(\Phi)$ in terms of Fourier transforms of $\phi \in \Phi$ was given by de Boor, DeVore, and Ron [2]:

$$\mathcal{S}^k(\Phi) = \left\{ f \in L_2 \mid \widehat{f} = \sum_{\varphi \in \Phi} \tau_\varphi \widehat{\varphi}, \ \tau_\varphi \ \text{ is } \ 2^{k+1}\pi\text{-periodic}, \ \varphi \in \Phi \right\}, \tag{2.1}$$

where $\widehat{f}$ is the Fourier transform of $f \in L_2$,

$$\widehat{f}(\omega) := \int_{-\infty}^{\infty} f(x) e^{-ix\omega} dx.$$

In [3] the bracket product of two functions $f, g \in L_2$ is defined by

$$[f, g](\omega) := \sum_{\alpha \in \mathbb{Z}} \widehat{f}(\omega + 2\alpha\pi) \overline{\widehat{g}}(\omega + 2\alpha\pi).$$

It was first introduced by Jia and Micchelli [14] with some decay conditions on f and g. We note that the series converges absolutely for almost every $\omega \in \mathbb{R}$. Thus $[f, g]$ is well defined almost everywhere (a.e.). Moreover $[f, g] \in L_1[0, 2\pi]$. The Gramian matrix of a finite set $\Phi \subseteq L_2$ is given by $G(\Phi) := ([\varphi, \psi])_{\varphi, \psi \in \Phi}$. It is a nonnegative definite matrix for almost every ω.

Following [2], we say that Φ provides a quasi basis for $\mathcal{S}^0(\Phi)$ (that is, their integer shifts is a quasi basis) if

$$\det G(\Phi)(\omega) > 0 \quad \text{a.e. } \omega \in \sigma(\Phi) := \{\omega | \text{ rank } G(\Phi)(\omega) > 0\}.$$

The shifts of Φ are quasi-stable if there are two positive constants A and B such that

$$BI \leq G(\Phi) \leq AI \quad \text{a.e. } \omega \in \sigma(\Phi), \tag{2.2}$$

where I is the identity matrix of the same order as $G(\Phi)$ and, as usual, $M_1 \leq M_2$ for two $n \times n$ nonnegative definite matrices M_1 and M_2 means that for any $X = (x_1, \ldots, x_n) \in \mathbb{C}^n$ with X^* its complex conjugate,

$$XM_1X^* \leq XM_2X^*.$$

The bounds of a quasi-stable basis are the smallest A and largest B such that the inequalities in (2.2) hold.

The concepts of quasi basis and quasi-stable basis apply to $\mathcal{S}^k(\Phi), k \neq 0$, by scaling. We define the scaling operator sc on L_2 by $sc : f \to 2^{1/2}f(2\cdot), f \in L_2$. If Φ is a subset of L_2 and k is an integer, let $sc^k\Phi = \{\, sc^k f \,|\, f \in \Phi\}$. It is easily seen that $\mathcal{S}^k(\Phi) = sc^k\mathcal{S}^0(sc^{-k}\Phi)$ for any finite set Φ of L_2. The 2^{-k} shifts $\{\varphi(\cdot - 2^{-k}\alpha)|\alpha \in \mathbb{Z}, \varphi \in \Phi\}$ of Φ are said to form a quasi basis (quasi-stable basis, respectively) for $\mathcal{S}^k(\Phi)$ if and only if the integer shifts of $sc^{-k}\Phi$ form a quasi basis (quasi-stable basis, respectively) for $\mathcal{S}^0(sc^{-k}\Phi)$. In the case that the 2^{-k} shifts of Φ form a quasi-stable basis for $\mathcal{S}^k(\Phi)$, its bounds are defined to be those of the quasi-stable basis $\{\, (sc^{-k}\varphi)(\cdot - \alpha) \mid \varphi \in \Phi,\ \alpha \in \mathbb{Z}\}$ for $\mathcal{S}^0(sc^{-k}\Phi)$.

We recall that $\{\varphi(\cdot - \alpha)|\alpha \in \mathbb{Z}, \varphi \in \Phi\}$ is a Riesz basis with bounds A and B for $\mathcal{S}^0(\Phi)$ if and only if A is the smallest constant and B is the largest constant such that

$$BI \leq G(\Phi) \leq AI \quad \text{a.e. } \omega.$$

Therefore a Riesz basis $\{\varphi(\cdot - \alpha)|\alpha \in \mathbb{Z}, \varphi \in \Phi\}$ is a quasi-stable basis with $\sigma(\Phi) = \mathbb{R}$.

We begin our study with a characterization of when $\{\varphi(\cdot - \alpha)|\alpha \in \mathbb{Z}, \varphi \in \Phi\}$ is a frame for $\mathcal{S}^0(\Phi)$.

THEOREM 2.1. *Let $\Phi = \{\varphi_k\}_{k=1}^n \subseteq L_2$. Then for arbitrary nonnegative constants A and B, the following conditions are equivalent:*

$$B||f||^2 \leq \sum_{k=1}^n \sum_{\alpha\in\mathbb{Z}} |\langle f, \varphi_{k\alpha}\rangle|^2 \leq A||f||^2 \quad \forall f \in \mathcal{S}^0(\Phi), \tag{2.3}$$

$$BG(\Phi) \leq G^2(\Phi) \leq AG(\Phi) \quad \text{a.e.} \tag{2.4}$$

Proof. For any $f \in L_2$ we have by Plancherel's theorem

$$\sum_{k=1}^n \sum_{\alpha\in\mathbb{Z}} |\langle f, \varphi_{k\alpha}\rangle|^2 = \frac{1}{2\pi} \sum_{k=1}^n \int_0^{2\pi} |[f, \varphi_k]|^2, \tag{2.5}$$

(cf. [12, p. 67]), where $\varphi_\alpha(\cdot) = \varphi(\cdot - \alpha)$ for any $\varphi \in L_2$ and $\alpha \in \mathbb{Z}$.

For $f \in \mathcal{S}^0(\Phi)$, by (2.1), we can find a 2π-periodic mapping $m(\omega) = (m_1(\omega), \ldots, m_n(\omega)), \mathbb{R} \to \mathbb{C}^n$ such that

$$\widehat{f} = \sum_{l=1}^n m_l \widehat{\varphi}_l, \tag{2.6}$$

which yields

$$[f, \varphi_k] = \sum_{l=1}^{n} m_l [\varphi_l, \varphi_k], \quad k = 1, \dots, n.$$

Therefore it follows from (2.5) that

$$\sum_{k=1}^{n} \sum_{\alpha \in \mathbb{Z}} |\langle f, \varphi_{k\alpha} \rangle|^2 = \frac{1}{2\pi} \sum_{k=1}^{n} \int_0^{2\pi} \left| \sum_{l=1}^{n} m_l [\varphi_l, \varphi_k] \right|^2$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \sum_{l,j=1}^{n} m_l \overline{m}_j \sum_{k=1}^{n} [\varphi_l, \varphi_k] \overline{[\varphi_j, \varphi_k]} = \frac{1}{2\pi} \int_0^{2\pi} m G^2(\Phi) m^*.$$

On the other hand, it follows from Plancherel's theorem and (2.6) that

$$\|f\|^2 = \frac{1}{2\pi} \int_0^{2\pi} [f, f] = \frac{1}{2\pi} \int_0^{2\pi} m G(\Phi) m^*. \tag{2.7}$$

Thus we conclude that (2.3) is equivalent to

$$B \int_0^{2\pi} m G(\Phi) m^* \le \int_0^{2\pi} m G^2(\Phi) m^* \le A \int_0^{2\pi} m G(\Phi) m^* \tag{2.8}$$

for any 2π-periodic mapping m with

$$m G(\Phi) m^* \in L_1[0, 2\pi]. \tag{2.9}$$

For any (Lebesgue) measurable set $E \subseteq [0, 2\pi]$, replacing m in (2.8) by a 2π-periodic mapping that is equal to $m\chi_E$ for $\omega \in [0, 2\pi]$, we see that (2.8) is indeed equivalent to

$$B \int_E m G(\Phi) m^* \le \int_E m G^2(\Phi) m^* \le A \int_E m G(\Phi) m^*,$$

where χ_E is, as usual, the characteristic function of E. Lebesgue's theorem about Lebesgue point yields that

$$B m G(\Phi) m^* \le m G^2(\Phi) m^* \le A m G(\Phi) m^* \quad \text{a.e.}$$

for any m satisfying (2.9).

Since the entries of $G(\Phi)$ are all in $L_1[0, 2\pi]$, any 2π-periodic mapping m with $\|m_k\|_\infty < \infty, 1 \le k \le n$, satisfies (2.9). Therefore (2.4) and (2.8) are equivalent. The proof is complete.

Remark 2.2. After completing the first version of the manuscript, the author became aware that Ron and Shen [20] had established Theorem 2.1 by a different method.

If $\Phi = \{\varphi_k\}_{k=1}^n$ provides a quasi basis for $\mathcal{S}^0(\Phi)$, then $G(\Phi)(\omega)$ is invertible for $\omega \in \sigma(\Phi)$. Therefore (2.4) is equivalent to

$$B \le G(\Phi) \le A \quad \text{a.e. on } \sigma(\Phi).$$

Therefore we have proved the following known result.

Corollary 2.3. *Φ provides a quasi-stable basis for $\mathcal{S}^0(\Phi)$ if and only if Φ provides a quasi basis for $\mathcal{S}^0(\Phi)$ and $\{\varphi_k(\cdot - \alpha) \mid \alpha \in \mathbb{Z}, 1 \le k \le n\}$ is a frame for $\mathcal{S}^0(\Phi)$.*

3. The splitting trick in shift-invariant spaces. Now we establish some inequalities about the splitting trick. These inequalities are essential to the construction of wavelet frame packets.

LEMMA 3.1. *Assume $\Phi = \{\varphi_k\}_{k=1}^n \subseteq L_2$. Let $\Psi = \{\psi_j\}_{j=1}^n \subseteq \mathcal{S}^0(\Phi)$ be given by*

$$\widehat{\psi}_j = \sum_{k=1}^n P_{jk}\widehat{\varphi}_k, \quad j = 1, \dots, n, \tag{3.1}$$

where P_{jk} are 2π-periodic functions, $j, k = 1, \dots, n$. If for $P = (P_{jk})_{j,k=1}^n$, there are some constants C' and C such that

$$C'I \leq P^*P \leq CI \quad \text{a.e. on } \sigma(\Phi), \tag{3.2}$$

then

$$\begin{aligned} &C' \sum_{k=1}^n \sum_{\alpha\in\mathbb{Z}} |\langle f, \varphi_{k\alpha}\rangle|^2 \leq \sum_{j=1}^n \sum_{\alpha\in\mathbb{Z}} |\langle f, \psi_{j\alpha}\rangle|^2 \\ &\leq C \sum_{k=1}^n \sum_{\alpha\in\mathbb{Z}} |\langle f, \varphi_{k\alpha}\rangle|^2 \quad \forall f \in L_2, \end{aligned} \tag{3.3}$$

where, as before, $\varphi_{k\alpha} = \varphi_k(\cdot - \alpha)$ and $\psi_{j\alpha} = \psi_j(\cdot - \alpha)$.

Moreover, if Φ provides a quasi-stable basis, (3.2) *and* (3.3) *are equivalent.*

Proof. For any $f \in L_2$ we have by (3.1) that

$$[f, \psi_j] = \sum_{k=1}^n \overline{P}_{jk}[f, \varphi_k], \quad j = 1, \dots, n.$$

Consequently

$$\begin{aligned} \sum_{j=1}^n |[f, \psi_j]|^2 &= \sum_{k,k'=1}^n \left(\sum_{j=1}^n \overline{P}_{jk} P_{jk'} \right) [f, \varphi_k]\overline{[f, \varphi_{k'}]} \\ &= XP^*PX^*, \end{aligned}$$

where $X = X(\omega) = ([f, \varphi_1], \dots, [f, \varphi_n])$.

It follows from (2.5) that (3.3) is equivalent to the inequalities

$$C' \int_0^{2\pi} XX^* \leq \int_0^{2\pi} XP^*PX^* \leq C \int_0^{2\pi} XX^* \quad \forall f \in L_2. \tag{3.4}$$

For any measurable set $E \subseteq [0, 2\pi]$ and $f \in L_2$ we define a function g by letting

$$\widehat{g}(\omega) = \widehat{f}(\omega) \sum_{\alpha\in\mathbb{Z}} \chi_E(\omega + \alpha).$$

Then $g \in L_2$ and $[g, \varphi_k] = [f, \varphi_k]\chi_E$. Since the measurable sets E are arbitrary we observe, as in the proof of Theorem 2.1, that (3.4) is equivalent to

$$C'XX^* \leq XP^*PX^* \leq CXX^* \text{ a.e.}$$

Since $X(\omega) = 0$ for any $f \in L_2$ and $\omega \in [0, 2\pi]\backslash\sigma(\Phi)$, we actually have proved that (3.3) holds if and only if

$$C'XX^* \leq XP^*PX^* \leq CXX^* \quad \text{a.e.} \quad \text{on } \sigma(\Phi) \tag{3.5}$$

for any $X = ([f, \varphi_1], \ldots, [f, \varphi_n])$ with $f \in L_2$. On the other hand, (3.2) means that (3.5) holds for any $X \in \mathbb{C}^n$. Therefore (3.2) implies (3.3).

Assume further that Φ provides a quasi-stable basis. We claim that for any 2π-periodic and measurable mapping $X = X(\omega) = (x_1(\omega), \ldots, x_n(\omega))$ with $X(\omega) = 0$ for $\omega \in [0, 2\pi]\backslash\sigma(\Phi)$ and $||x_k||_\infty < \infty, 1 \leq k \leq n$, there is some $f \in L_2$ satisfying

$$X = ([f, \varphi_1], \ldots, [f, \varphi_n]).$$

Indeed, let f be defined by (2.6) with $m = XG^{-1}(\Phi)$ for $\omega \in \sigma(\Phi)$ and $m = 0$ for $\omega \in [0, 2\pi]\backslash\sigma(\Phi)$. Since

$$mG(\Phi)m^* = XG^{-1}(\Phi)X^*,\ \omega \in \sigma(\Phi),$$

and all entries of $G^{-1}(\Phi)$ belong to $L_\infty(\sigma(\Phi))$, we have $mGm^* \in L_1(\sigma(\Phi))$. Therefore $f \in \mathcal{S}^0(\Phi)$ and f satisfies the required equality.

Since, for any fixed $\omega \in \sigma(\Phi), X$ in (3.5) may range over the whole $\mathbb{C}^n$, (3.2) is equivalent to (3.3). The proof is complete.

THEOREM 3.2 (splitting trick). *Let $\varphi \in L_2$ and $\{\psi_j\}_{j=0}^1 \subseteq \mathcal{S}^1(\varphi)$ be given by*

$$\widehat{\psi}_j(\omega) = \sqrt{2}m_j(\omega/2)\widehat{\varphi}(\omega), \quad j = 0, 1, \tag{3.6}$$

for some 2π-periodic $m_j, j = 0, 1$. Define the matrix $M(\omega) := (m_j(\omega + k\pi))_{j,k=0}^1$. Then the condition that

$$C'I \leq M^*(\omega/2)M(\omega/2) \leq CI \quad \textit{a.e. on } \sigma(\Phi) \tag{3.7}$$

implies that for any $f \in L_2$

$$\begin{aligned} &C' \sum_{\alpha\in\mathbb{Z}} |\langle f, \varphi(\cdot - \alpha/2)\rangle|^2 \\ &\leq \sum_{j=0,1} \sum_{\alpha\in\mathbb{Z}} |\langle f, \psi_{j\alpha}\rangle|^2 \\ &\leq C \sum_{\alpha\in\mathbb{Z}} |\langle f, \varphi(\cdot - \alpha/2)\rangle|^2. \end{aligned} \tag{3.8}$$

Moreover, if $\{\varphi(\cdot - \alpha/2)\}_{\alpha\in\mathbb{Z}}$ is a quasi-stable basis, then (3.7) *and* (3.8) *are equivalent.*

Proof. Define $\varphi_k(\cdot) = \varphi(\cdot - k/2), k = 0, 1$. Then $\mathcal{S}^1(\varphi) = \mathcal{S}^0(\Phi)$, where $\Phi = \{\varphi_k\}_{k=0}^1$, and (3.6) may be represented as (3.1) for $n = 2$ and some 2π-periodic $P_{jk}, j, k = 0, 1$.

Recall that the matrix $P = (P_{jk})_{jk=0}^1$ is associated with M by the equation (cf. [17] and [18, p. 86]),

$$M(\omega) = P(2\omega)U(\omega) \tag{3.9}$$

with the matrix

$$U(\omega) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ e^{-i\omega} & -e^{-i\omega} \end{pmatrix}$$

being 2π-periodic and unitary for any ω. Therefore

$$M^*(\omega/2)M(\omega/2) = U^*(\omega/2)P^*(\omega)P(\omega)U(\omega/2) \quad \forall \omega \in \mathbb{R}.$$

The theorem follows immediately from Lemma 3.1 and the equivalence of (3.2) and (3.7).

COROLLARY 3.3. *Suppose that the conditions of Theorem* 3.2 *are satisfied. Assume furthermore that* $0 < C' \leq C < \infty$ *and* $\{\varphi(\cdot - \alpha/2) \mid \alpha \in \mathbb{Z}\}$ *is a frame (quasi-stable basis, Riesz basis, respectively) of* $\mathcal{S}^1(\varphi)$ *with bounds* A *and* B. *Then* $\{\psi_j(\cdot - \alpha) \mid 0 \leq j \leq 1,\ \alpha \in \mathbb{Z}\}$ *is also a frame (quasi-stable basis, Riesz basis, respectively) of* $\mathcal{S}^1(\varphi)$ *with bounds* $A' \leq CA$ *and* $B' \geq C'B$.

Proof. The conclusion in the frame case is obvious by Theorem 3.2. As for the quasi-stable basis case, we note that $\mathcal{S}^0(\Psi) = \mathcal{S}^0(\Phi)$, where $\Phi = \{\varphi_k\}_{k=0}^1$ is as given in the proof of Theorem 3.2 and $\Psi = \{\psi_k\}_{k=0}^1$. It follows from the equality $G(\Psi) = PG(\Phi)P^*$ that $\text{rank}G(\Psi) = \text{rank}G(\Phi)$ for any ω. Moreover, by assumption, Φ provides a quasi-stable basis of $\mathcal{S}^0(\Phi)$ with bounds A and B. Since we already know that $\{\psi_j(\cdot - \alpha) \mid 0 \leq j \leq 1,\ \alpha \in \mathbb{Z}\}$ is also a frame of $\mathcal{S}^0(\Phi)$ with bounds $A' \leq CA$ and $B' \geq C'B$, the conclusion follows from Corollary 2.3.

Remark 3.4. The analogue of Theorem 3.2 and Corollary 3.3 hold when $\mathcal{S}^1(\varphi)$ and $m_j(2^{-1}\omega)$ are replaced by $\mathcal{S}^{j_0}(\varphi)$ and $m_j(2^{-j_0}\omega)$, respectively, where $j_0 \in \mathbb{Z}$ is any integer.

In what follows, we need to apply several splitting steps. To this end it is useful to introduce some index sets. Let $\mathcal{M} = \bigcup_{k=1}^{\infty}\{0,1\}^k$ and $\mathcal{M}_L = \bigcup_{k=1}^{L}\{0,1\}^k, L < \infty$. For any $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_k) \in \mathcal{M}$ we set $|\varepsilon| = k$ and $l_\varepsilon = \sum_{j=1}^{k} 2^{-j}\varepsilon_j$. Associated with ε is a dyadic interval $I_\varepsilon = [l_\varepsilon, l_\varepsilon + 2^{-|\varepsilon|})$.

A subset $\mathcal{N} \subseteq \mathcal{M}$ is called an admissible set if $L_\varepsilon = \{I_\varepsilon\}, \varepsilon \in \mathcal{N}$, partition the unit interval [0,1). This means that [0,1) is the union of $I_\varepsilon, \varepsilon \in \mathcal{N}$, any of which does not intersect the others. Both of the following $\mathcal{N}$ are admissible sets: $\mathcal{N} = \{\varepsilon \,|\, |\varepsilon| = k\}$, where $k > 0$ fixed; $\mathcal{N} = \{(0), (1,0), (1,1,0), (1,1,1,0), (1,1,1,1,0), \ldots\}$.

LEMMA 3.5. *The following statements are true:*

(i) *Given two dyadic intervals,* I_ε *and* I_τ, *either their intersection is empty or one contains the other.*

(ii) *Let* $\mathcal{N}$ *be an admissible set. For any* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_k) \in \mathcal{N}$ *and any* $\varepsilon_{k+1}, \ldots, \varepsilon_j$,

$$\varepsilon' := (\varepsilon_1, \ldots, \varepsilon_k, \varepsilon_{k+1}, \ldots, \varepsilon_j) \notin \mathcal{N},$$

where $k < j$.

(iii) *If* $\mathcal{N} \subseteq \mathcal{M}_L$ *is an admissible set and* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_L) \in \mathcal{N}$, *then*

$$\varepsilon^0 := (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{L-1}, 1 - \varepsilon_L) \in \mathcal{N}.$$

Proof. To prove (i) we assume that $I_\varepsilon \cap I_\tau$ is not empty and, without loss of generality, assume $|\tau| \leq |\varepsilon|$. It follows from the definition that $I_\tau \subseteq I_\varepsilon$. Moreover $I_\tau = I_\varepsilon$ if and only if $|\tau| = |\varepsilon|$.

The verification of (ii) is easy by noting that $I_{\varepsilon'} \subseteq I_\varepsilon$.

As for the proof of (iii), we observe that $\mathcal{N}$ contains no $\tau \neq \varepsilon^0$ such that $I_{\varepsilon^0} \subseteq I_\tau$. For, otherwise, we have also $I_\varepsilon \subseteq I_\tau$, a contradiction. Note that $\mathcal{N} \subseteq \mathcal{M}_L$ is admissible. ε^0 must belong to $\mathcal{N}$. The proof is complete.

Given $\varphi \in L_2$ and 2π-periodic functions $m_j, j = 0, 1$, we apply the splitting trick, as in Theorem 3.2, to $\mathcal{S}^{j_0}(\varphi)$ and their resulting subspaces successively, where $j_0 \in \mathbb{Z}$

fixed, getting a sequence of resulting functions φ_ε (we adopt the notation φ_ε instead of ψ's as in Theorem 3.2) given by their Fourier transform

$$\widehat{\varphi}_\varepsilon(\omega) = 2^{|\varepsilon|/2}\widehat{\varphi}(\omega)\prod_{j=1}^{k} m_{\varepsilon_j}(2^{j-1-j_0}\omega), \tag{3.10}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_k)$. However, some conditions to ensure $\varphi_\varepsilon \in L_2$, and hence $\varphi_\varepsilon \in \mathcal{S}^{j_0}(\varphi)$ should be imposed on $m_j, j = 0, 1$. The following lemma gives such conditions. The proof is easy and thus omitted.

LEMMA 3.6. *Let $\varphi \in L_2$ and φ_ε be as given by* (3.10). *If the measurable and 2π-periodic functions m_j satisfy $\|m_j\|_\infty < \infty, j = 0, 1,$ then $\varphi_\varepsilon \in \mathcal{S}^{j_0}(\varphi)$.*

Remark 3.7. Let c^ε be the sequence of coefficients $\{c^\varepsilon_\alpha\}_{\alpha\in\mathbb{Z}}$, where

$$c^\varepsilon_\alpha := \langle f, \varphi_\varepsilon(\cdot - 2^{|\varepsilon|-j_0}\alpha)\rangle, \quad \alpha \in \mathbb{Z}.$$

It is worth pointing out the well-known fact that the coefficients can be computed by Mallat's algorithm successively whenever the coefficients $\langle f, \varphi(\cdot - 2^{-j_0}\alpha)\rangle$, $\alpha \in \mathbb{Z}$, are given. This is due to the relation between φ_ε and φ. In fact, if we set $m_j(\omega) := \sum_{\beta\in\mathbb{Z}} h_{j\beta}e^{-i\beta\omega}, j = 0, 1,$ then

$$c^{(\varepsilon_1,\dots,\varepsilon_k)}_\alpha = F_{\varepsilon_k}c^{(\varepsilon_1,\dots,\varepsilon_{k-1})}(\alpha), \quad \alpha \in \mathbb{Z},$$

where $F_j, j = 0, 1,$ is an operator from $l_2(\mathbb{Z})$ into $l_2(\mathbb{Z})$,

$$F_j s(\alpha) = \sqrt{2}\sum_\beta s_\beta \overline{h}_{j\beta-2\alpha}, \quad s \in l_2(\mathbb{Z}).$$

Let $m_j, j = 0, 1$, satisfy the condition in Lemma 3.6. Denote by $\Lambda(\omega)$ and $\lambda(\omega)$, for a.e. $\omega \in \mathbb{R}$, the maximal and minimal eigenvalues, respectively, of the matrix $M^*(\omega/2)M(\omega/2)$. Then the functions $\Lambda(\omega)$ and $\lambda(\omega)$ are in fact 2π-periodic by (3.9). Moreover, they are measurable. Let $\Lambda = \|\Lambda(\omega)\|_\infty$ and $\lambda = \|\lambda(\omega)\|_\infty$. Then

$$\lambda I \le M^*M \le \Lambda I, \quad \text{a.e.} \tag{3.11}$$

Now we can generalize Theorem 3.2 as follows.

THEOREM 3.8. *Let Λ and λ be defined as above. Assume $\varphi \in L_2$ and define φ_ε as in* (3.10). *Then for any admissible set $\mathcal{N} \subseteq \mathcal{M}_L$, $L < \infty$ fixed, we have for any $f \in L_2$ and $j_0 \in \mathbb{Z}$*

$$\begin{aligned} &B_L\sum_{\alpha\in\mathbb{Z}}|\langle f, \varphi(\cdot - 2^{-j_0}\alpha)\rangle|^2 \\ &\le \sum_{\varepsilon\in\mathcal{N}}\sum_{\alpha\in\mathbb{Z}}|\langle f, \varphi_\varepsilon(\cdot - 2^{|\varepsilon|-j_0}\alpha)\rangle|^2 \\ &\le A_L\sum_{\alpha\in\mathbb{Z}}|\langle f, \varphi(\cdot - 2^{-j_0}\alpha)\rangle|^2, \end{aligned} \tag{3.12}$$

where $A_L = \max\{1, \Lambda^L\}$ and $B_L = \min\{1, \lambda^L\}$.

Consequently, if $M(\omega)$ is unitary almost everywhere, then for any admissible set $\mathcal{N} \subseteq \mathcal{M}$ (not necessarily $\mathcal{N} \subseteq \mathcal{M}_L$) we have for $f \in L_2$

$$\sum_{\varepsilon\in\mathcal{N}}\sum_{\alpha\in\mathbb{Z}}|\langle f, \varphi_\varepsilon(\cdot - 2^{|\varepsilon|-j_0}\alpha)\rangle|^2 = \sum_{\alpha\in Z}|\langle f, \varphi(\cdot - 2^{-j_0}\alpha)|^2. \tag{3.13}$$

Proof. We observe first that if $M(\omega)$ is unitary for almost every ω, then $\Lambda = \lambda = 1$. Therefore the equality (3.13) follows from (3.12). It remains to prove (3.12).

To prove (3.12) we work by induction. When $L = 1$, $A_L \geq \Lambda$ and $B_L \leq \lambda$. Therefore (3.12) follows from (3.8) and (3.11).

Suppose that (3.12) holds for $L = L_0$. Assume that $\mathcal{N} \subset \mathcal{M}_{L_0+1}$ is an admissible set. We divide it into two parts, $\mathcal{N} = \mathcal{N}_{L_0+1} \cup (\mathcal{N} \cap \mathcal{M}_{L_0})$, where

$$\mathcal{N}_L = \{\, \varepsilon \,|\, \varepsilon \in \mathcal{N} \quad \text{and} \quad |\varepsilon| = L\}.$$

For any $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{L_0+1}) \in \mathcal{N}_{L_0+1}$, it follows from Lemma 3.5 that

$$\varepsilon^0 = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{L_0}, 1 - \varepsilon_{L_0+1}) \in \mathcal{N}_{L_0+1}.$$

Associated with any pair of $\varepsilon, \varepsilon_0 \in \mathcal{N}_{L_0+1}$, we set

$$\tau = (\varepsilon_1, \ldots, \varepsilon_{L_0}). \tag{3.14}$$

Appealing to the analogy of Theorem 3.2 (see Remark 3.4) we obtain for $f \in L_2$

$$\begin{aligned}
&\lambda \sum_{\alpha \in \mathbb{Z}} |\langle f, \varphi_\tau(\cdot - 2^{L_0 - j_0}\alpha)\rangle|^2 \\
&\leq \sum_{\alpha \in \mathbb{Z}} \big(|\langle f, \varphi_\varepsilon(\cdot - 2^{L_0+1-j_0}\alpha)\rangle|^2 + |\langle f, \varphi_{\varepsilon^0}(\cdot - 2^{L_0+1-j_0}\alpha)\rangle|^2 \big) \\
&\leq \Lambda \sum_{\alpha \in \mathbb{Z}} |\langle f, \varphi_\tau(\cdot - 2^{L_0 - j_0}\alpha)\rangle|^2 .
\end{aligned} \tag{3.15}$$

Since $\mathcal{N}' = \mathcal{N}_{L_0} \cup \{\tau\}_{\varepsilon \in \mathcal{N}_{L_0+1}} \subseteq \mathcal{M}_{L_0}$ is an admissible set, where τ is associated with ε by (3.14) for any $\varepsilon \in \mathcal{N}_{L_0+1}$, it follows from the induction assumption that for $f \in L_2$

$$\begin{aligned}
&B_{L_0} \sum_{\alpha \in \mathbb{Z}} |\langle f, \varphi(\cdot - 2^{-j_0}\alpha)|^2 \\
&\leq \sum_{\alpha \in \mathbb{Z}} \Big(\sum_{\delta \in \mathcal{N} \cap \mathcal{M}_{L_0}} |\langle f, \varphi_\delta(\cdot - 2^{|\delta| - j_0}\alpha)\rangle|^2 + \sum_{\tau} |\langle f, \varphi_\tau(\cdot - 2^{L_0 - j_0}\alpha)\rangle|^2 \Big) \\
&\leq A_{L_0} \sum_{\alpha \in \mathbb{Z}} |\langle f, \varphi(\cdot - 2^{-j_0}\alpha)\rangle|^2 .
\end{aligned}$$

Now (3.15) implies (3.12) for $L = L_0 + 1$. The proof is complete.

It is very important that A_L and B_L in (3.12) are independent of j_0.

Corollary 3.9. *In addition to the conditions of Theorem* 3.2 *we assume that* $0 < \lambda \leq \Lambda < \infty$ *and* $\{\varphi(\cdot - 2^{-j_0}\alpha) \mid \alpha \in \mathbb{Z}\}$ *is a frame (Riesz basis, respectively) of* $\mathcal{S}^{j_0}(\varphi)$ *with bounds A and B. Then for any admissible set* $\mathcal{N} \subseteq \mathcal{M}_L$*, where* $L < \infty$ *is fixed, the sequence*

$$\{\varphi_\varepsilon(\cdot - 2^{|\varepsilon| - j_0}\alpha) | \alpha \in \mathbb{Z}, \varepsilon \in \mathcal{N}\} \tag{3.16}$$

is a frame (Riesz basis, respectively) of $\mathcal{S}^{j_0}(\varphi)$ *with bounds* A' *and* B' *with* $A' \leq AA_L$ *and* $B' \geq BB_L$*.*

Consequently, (3.16) *is a frame (Riesz basis, respectively) with bounds A and B for any admissible set $\mathcal{N}$ when M is unitary everywhere.*

Proof. The conclusions about the frame case follow easily from a special case of Theorem 3.8, in which (3.12) need hold only for any $f \in \mathcal{S}^{j_0}(\varphi)$.

To prove the result about the Riesz basis case, we need only to verify that the sequence in (3.16) is a Riesz basis. Since $\{\, \varphi(\cdot - 2^{-j_0}\alpha) \mid \alpha \in \mathbb{Z} \,\}$ is a Riesz basis of $\mathcal{S}^{j_0}(\phi)$, by applying the analogy of Corollary 3.3 for $\mathcal{S}^{j_0}(\varphi)$ (see Remark 3.4) and the resulting subspaces successively, we conclude that the sequence (3.16) is indeed a Riesz basis. The proof is complete.

4. Wavelet frame packets. In this section we construct the wavelet frame packets for L_2. The general theory is illustrated with an example of compactly supported tight wavelet frame packets.

THEOREM 4.1. *Let S be a subset of $\mathbb{Z}^2$ and $\{\psi_k(\cdot - 2^{-j}\alpha)|\alpha \in \mathbb{Z}, (k,j) \in S\}$ a frame (Riesz basis, respectively) with bounds A and B for L_2. Assume that $\mathcal{N}(k,j) \subset \mathcal{M}_L$ is an admissible set for any $(k,j) \in S$, where $L < \infty$ is independent of $(k,j) \in S$. Let m_0 and m_1 satisfy the conditions of Lemma* 3.6, *and the associated matrix $M(\omega) = ((m_j(\omega + k\pi))_{j,k=0}^1$ satisfies* (3.11) *for some λ and Λ with $0 < \lambda \le \Lambda < \infty$. Moreover, let $\psi_{k\varepsilon}$ be as given in* (3.10) *by replacing φ with ψ_k, and let A_L and B_L be as given in Theorem* 3.8. *Then the sequence*

$$\{\psi_{k\varepsilon}(\cdot - 2^{|\varepsilon|-j}\alpha)|\alpha \in \mathbb{Z}, \varepsilon \in \mathcal{N}(k,j) \text{ and } (k,j) \in S\} \tag{4.1}$$

is a frame (Riesz basis, respectively) of L^2 with bounds $A' \le AA_L$ and $B' \ge BB_L$.

Remark 4.2. If $\mathcal{N}(k,j)$ is empty for some $(k,j) \in S$, we take no splitting on the corresponding sequence $\{\psi_k(\cdot - 2^{-j}\alpha)|\alpha \in \mathbb{Z}\}$.

The proof of Theorem 4.1 is the same as that of Corollary 3.9 in section 3 only with the difference in that we should use the inequality (3.12) for all $f \in L_2$. We omit it.

Since there is a large variety of choices of $\mathcal{N}(k,j)$ for any $(k,j) \in S$, we can pick many frames from (4.1). Combining the terminologies of wavelet frames and wavelet packets, we call all these frames the wavelet frame packets.

Remark 4.3. It is easily seen that A_L and B_L^{-1} defined in Theorem 3.8 grow at most exponentially for large L. Therefore, the estimation for the bounds of the resulting frames (Riesz bases, respectively) in Theorem 4.1 is sharper than that in [8].

On the other hand, if the matrix $M(\omega)$ is unitary for any ω, we have $A_L = B_L = 1$ for any L. It implies that the bounds of the resulting frames (Riesz bases, respectively) are unchanged. These bounds are optimal. In this case, the estimation of [6] cannot give the optimal bounds.

As stated in Remark 3.7, the coefficients of a function f with respect to sequence in (4.1) may be computed by Mallat's algorithm. It is possible to find a best frame for a given f. The algorithm is the same as in wavelet packets [10].

Example 4.4 (tight wavelet frame packets). Let m_0 be a 2π-periodic trigonometric such that

$$m_0(0) = 1, \quad |m_0(\omega)|^2 + |m_0(\omega + \pi)|^2 = 1.$$

Define φ and ψ by $\widehat{\varphi}(\omega) = \prod_{k=1}^{\infty} m_0(2^{-k}\omega)$ and $\widehat{\psi}(\omega) = m_1(\omega/2)\widehat{\varphi}(\omega/2)$, respectively, where as usual, $m_1(\omega) = e^{-i\omega}\overline{m}_0(\omega + \pi)$. It is easy to see that the matrix $M(\omega) = (m_j(\omega + k\pi))_{j,k=0}^1$ is unitary for any ω.

It was proved in [16] that $\{2^{k/2}\psi(2^k \cdot - \alpha)\}_{k,\alpha\in\mathbb{Z}}$ is a tight frame for L_2 with bound 1. This corresponds to the case $\psi_k(\cdot) := 2^{k/2}\psi(2^k \cdot)$ and $S = \{\,(k,k) \mid k \in \mathbb{Z}\,\}$ in (4.1).

We are more interested in the sequence

$$\{\varphi(\cdot-\alpha),\ \psi_k(\cdot-2^{-k}\alpha)\mid\alpha\in\mathbb{Z},k\geq 0\}. \tag{4.2}$$

From the proof of [12, Proposition 6.2.3] we see that the sequence (4.2) is also a tight frame of L_2 with bound 1.

As in (3.10), we define for $k>0$ and $\varepsilon=(\varepsilon_1,\dots,\varepsilon_k)$ the function $\psi_{k\varepsilon}$ by

$$\widehat{\psi}_{k\varepsilon}(\omega)=\widehat{\psi}(2^{-k}\omega)\prod_{j=1}^{k}m_{\varepsilon_j}(2^{-j}\omega)=\prod_{j=1}^{\infty}m_{\varepsilon_j}(2^{-j}\omega),$$

where $\varepsilon_{k+1}=1$ and $\varepsilon_j=0$ for $j\geq k+2$. The last equality follows from the relation between $\widehat{\psi}$ and $\widehat{\phi}$. If we split the 2^{-k} shifts of $\psi_k,k>0$, exactly k times at all possibilities, the resulting sequence is the set of integer shifts of all functions $\psi_{k\varepsilon}$ $\forall\varepsilon$ with $|\varepsilon|=k$. Therefore, taking (4.2) as the original frame, we get the following resulting frame:

$$\{\varphi(\cdot-\alpha),\psi(\cdot-\alpha),\psi_{k\varepsilon}(\cdot-\alpha)\mid\alpha\in\mathbb{Z},|\varepsilon|=k,k>0\}. \tag{4.3}$$

By the unitary property of the matrix M and Theorem 4.1, (4.3) is also a tight frame of L_2 with bound 1.

Adopting the notation of [10], we can check easily that (4.3) is nothing but the sequence

$$\{W_n(\cdot-\alpha)|n\geq 0,\alpha\in\mathbb{Z}\}, \tag{4.4}$$

where W_n is defined by its Fourier transform

$$\widehat{W_n}(\omega)=\prod_{j=1}^{\infty}m_{\varepsilon_j}(2^{-j}\omega),\quad n=\sum_{j=1}^{\infty}2^{j-1}\varepsilon_j. \tag{4.5}$$

Besides (4.4) we may obtain many tight frames of L_2 as follows. In fact, as in [10], for any partition $P=\{I_{kn}\}$ of the nonnegative integers of the form $I_{kn}=\{2^kn,\dots,2^k(n+1)-1\}$ we consider the set

$$\{2^{\frac{k}{2}}W_n(2^k\cdot-\alpha)|I_{kn}\in P,\alpha\in\mathbb{Z}\}. \tag{4.6}$$

We claim that (4.6) is a tight frame of L_2 with bound 1. Indeed, for any I_{nk}, we define $\varepsilon_j,j>k$, by $n=\sum_{j=k+1}^{\infty}2^{j-k-1}\varepsilon_j$ and the function $\widetilde{W}_{nk}(\cdot)=2^{\frac{k}{2}}W_n(2^k\cdot)$. It follows from (4.5) that

$$\widehat{\widetilde{W}}_{nk}(\omega)=2^{-k/2}\widehat{W}_n(2^{-k}\omega)=2^{-k/2}\prod_{j=1}^{\infty}m_{\varepsilon_{j+k}}(2^{-j-k}\omega).$$

For any $\varepsilon=(\varepsilon_1,\dots,\varepsilon_k)$ we define $\widetilde{W}_{nk\varepsilon}$ as in (3.10) by replacing φ with $\widetilde{W}_{nk}$. Then

$$\widehat{\widetilde{W}}_{nk\varepsilon}(\omega)=\prod_{j=1}^{\infty}m_{\varepsilon_j}(2^{-j}\omega).$$

Consequently,

$$\widetilde{W}_{nk\varepsilon}=W_{n'},\quad n'=2^kn+\sum_{j=1}^{k}2^{j-1}\varepsilon_j.$$

By assumption on $P = \{I_{kn}\}$ we have

$$\{\widetilde{W}_{nk\varepsilon} | |\varepsilon| = k, I_{kn} \in P\} = \{W_{n'} | n' \in \mathbb{Z}, n' \geq 0\}.$$

This means that the set $\{\widetilde{W}_{nk\varepsilon}(\cdot - \alpha) | |\varepsilon| = k, I_{kn} \in P$ and $\alpha \in \mathbb{Z}\}$ is just the sequence (4.4) and, therefore, a tight frame with bound 1 for L_2. On the other hand, the equality (3.13) yields that for any $f \in L_2$

$$\sum_{I_{kn} \in P} \sum_{\alpha \in \mathbb{Z}} |\langle f, \widetilde{W}_{nk}(\cdot - 2^{-k}\alpha)\rangle|^2 = \sum_{I_{kn} \in P} \sum_{|\varepsilon| = k} \sum_{\alpha \in \mathbb{Z}} |\langle f, \widetilde{W}_{nk\varepsilon}(\cdot - \alpha)\rangle|^2.$$

We conclude that $\{ \widetilde{W}_{nk}(\cdot - 2^{-k}\alpha) \mid I_{kn} \in P, \alpha \in \mathbb{Z} \}$ is also a tight frame with bound 1. Note that the last sequence is just (4.6). Therefore, for any $P = \{I_{nk}\}$ as above, the sequence (4.6) is a tight frame with bound 1, as claimed .

By formula (1.2) any $f \in L_2$ can be represented as a convergent series in L_2:

$$f(x) = \sum_{I_{kn} \in P} \sum_{\alpha} c_{\alpha}^{nk} 2^{k/2} W_n(2^k x - \alpha)$$

with coefficients $c_{\alpha}^{nk} = \langle f, 2^{k/2} W_n(2^k \cdot - \alpha)\rangle$.

Acknowledgments. The author thanks the late Prof. R. L. Long of the Institute of Mathematics, Chinese Academy of Sciences, for his valuable help and the referees for their thoughtful and valuable comments that improve the presentation of the paper significantly. The author also thanks Dr. Z. Shen for making available to him a copy of [20] before it appeared.

REFERENCES

[1] C. de Boor, R. DeVore, and A. Ron, *On the construction of multivariate (pre) wavelets*, Constr. Approx., 9 (1993), pp. 123–166.

[2] C. de Boor, R. DeVore, and A. Ron, *The structure of finitely generated shift-invariant subspaces*, J. Funct. Anal., 119 (1994), pp. 37–78.

[3] C. de Boor, R. DeVore, and A. Ron, *Approximation from shift-invariant subspaces in $L_2(\mathbb{R}^d)$*, Trans. Amer. Math. Soc., 341 (1994), pp. 787–806.

[4] D. R. Chen, *On the existence and constructions of orthonormal wavelets on $L_2(\mathbb{R}^s)$*, Proc. Amer. Math. Soc., 125 (1997), pp. 2883–2889.

[5] D. R. Chen, *On the constructions of pre-wavelets and Riesz wavelet basis in $L_2(\mathbb{R}^s)$*, Acta Math. Appl. Sinica, 14 (1998), pp. 129–133.

[6] C. K. Chui and C. Li, *Nonorthogonal wavelet packets*, SIAM J. Math. Anal., 24 (1993), pp. 712–738.

[7] C. K. Chui and J. Z. Wang, *On compactly supported spline wavelets and a duality principle*, Trans. Amer. Math. Soc., 330 (1992), pp. 903–912.

[8] A. Cohen and I. Daubechies, *On the instability of arbitrary biorthogonal wavelet packets*, SIAM J. Math. Anal., 24 (1993), pp. 1340–1354.

[9] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal wavelet bases and their related subband coding schemes*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[10] R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory, 3 (1992), pp. 713–718.

[11] I. Daubechies, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Trans. Inform. Theory, 36 (1990), pp. 961–1005.

[12] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[13] R. Q. Jia and C. A. Micchelli, *On linear independence for integer translates of a finite number of functions*, Proc. Edinburgh Math. Soc., 36 (1992), pp. 69–85.

[14] R. Q. Jia and C. A. Micchelli, *Using the refinement equations for the construction of pre-wavelets* II: *Power of two*, in Curves and Surfaces, P.-J. Laurent, A. Le Mehaute, and L. L. Schumaker, eds., Academic Press, New York, 1991, pp. 209–246.

[15] R. Q. Jia and Z. W. Shen, *Multiresolution and wavelets*, Proc. Edinburgh Math. Soc., 37 (1994), pp. 271–300.

[16] W. Lawton, *Tight frames of compactly supported affine wavelets*, J. Math. Phys., 31 (1990), pp. 1898–1901.

[17] R. L. Long and D. R. Chen, *Biorthogonal wavelets on* $\mathbb{R}^s$, Appl. Comput. Harmon. Anal., 2 (1995), pp. 230–242.

[18] Y. Meyer, *Wavelets and Operators* I, Cambridge University Press, London, 1992.

[19] S. D. Riemenschneider and Z. W. Shen, *Wavelets and prewavelets in low dimensions*, J. Approx. Theory, 71 (1992), pp. 18–38.

[20] A. Ron and Z. Shen, *Frames and stable bases for shift-invariant subspaces of* $L_2(\mathbb{R}^d)$, Canad. J. Math., 47 (1995), pp. 1051–1094.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 740–753

POINTWISE DECAY OF SOLUTIONS AND OF HIGHER DERIVATIVES TO NAVIER–STOKES EQUATIONS[*]

CHERIF AMROUCHE[†], VIVETTE GIRAULT[‡], MARIA ELENA SCHONBEK[§], AND TOMAS P. SCHONBEK[¶]

Abstract. In this paper we study the space-time asymptotic behavior of the solutions, and their derivatives, to the incompressible Navier–Stokes equations in dimension $2 \le n \le 5$. Using moment estimates we obtain that strong solutions to the Navier–Stokes equations which decay in L^2 at the rate of $\|u(t)\|_2 \le C(t+1)^{-\mu}$ will have the following pointwise space-time decay, for $0 \le k \le n/2$:

$$|D^\alpha u(x,t)| \le C_{k,m} \frac{1}{(t+1)^{\rho_0}(1+|x|^2)^{k/2}},$$

where $\rho_O = (1-2k/n)(m/2+\mu+n/4)$, $|\alpha| = m$ and $\mu > \frac{n}{4}$.

Key words. Navier–Stokes equations, derivatives, pointwise algebraic decay

AMS subject classifications. 35Q30, 76D05

PII. S0036141098346177

1. Introduction.

In this paper, we study the space-time decay of solutions to the incompressible Navier–Stokes equations in $\mathbb{R}^n$

$$\begin{aligned} u_t + u\cdot\nabla u + \nabla p &= \triangle u\,, \\ \operatorname{div} u &= 0\,, \\ u(x,0) = u_0(x) &\in \mathbf{X}\,, \end{aligned} \tag{1.1}$$

and of their derivatives. We assume $2 \le n \le 5$ and the space $\mathbf{X}$ will be specified below. Using moment techniques, we show that strong solutions and their derivatives of all orders decay pointwise at an algebraic rate as $|x| \to \infty$ and $t \to \infty$.

Questions of decay of solutions to the Navier–Stokes equations in different norms have been studied, among others, by Knightly [6], Kajikiya and Miyakawa [4], Kato [5], Kozono [7], Kozono and Ogawa [8], Schonbek [13], [10], Wiegner [18], and Zhang [20]. Of particular interest in the direction of the present paper are the results by S. Takahashi [17]. In this reference, Takahashi studies the pointwise decay of solutions with zero initial data to the Navier–Stokes equations with an external force, as well as the decay of the first derivatives of these solutions. Using a weighted-equation approach, he obtains pointwise decay rates both in time and space. The external force is assumed to decay at an algebraic rate in both space and time and the solutions are assumed bounded in some weighted $L^{q,s}$ norms, with $n/q + 2/s = 1$ and $q, s \in [2,\infty]$,

[*] Received by the editors October 26, 1998; accepted for publication (in revised form) July 28, 1999; published electronically March 21, 2000.
http://www.siam.org/journals/sima/31-4/34617.html

[†] Université de Pau et des Pays de l'Adour, France (Cherif.Amrouche@univ-pau.fr).

[‡] Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris, France (girault@emeu.ann.jussieu.fr).

[§] Department of Mathematics, University of California, Santa Cruz, CA 95064 (schonbek@math.ucsc.edu). This author was partially supported as Guest Professor Sofia Kowaleskaya, University of Kaiserlautern.

[¶] Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, FL 33431 (schonbek@acc.fau.edu).

(the limiting Serrin class), where $L^{q,s}$ denotes the space of all $u : \mathbb{R}^n \times (0, \infty) \to \mathbb{R}^n$ such that

$$\left\{ \int_0^\infty \left(\int_{\mathbf{R}^n} |u(x,t)|^q dx \right)^{s/q} dt \right\}^{1/s} < \infty.$$

Our results complement and extend Takahashi's results in the sense that in our case we have nonzero initial data but zero external force. Moreover, we are able to establish decay for derivatives of all orders. We note that since we are obtaining decay results for derivatives, we will work directly with strong solutions. These results can be derived for weak solutions provided we start at a sufficiently large time. Since in this case we are already in the regime where the solutions are smooth, we prefer to simplify notation and work directly with smooth solutions. The reader can also refer to [17], which presents a very detailed outline of what other authors in the field have done with related questions.

It is already clear at the level of the heat equation that there is a relation between the time decay and the space decay. This kind of balance will also be found for solutions to the Navier–Stokes equations. In particular the balance relation we obtain between the decay in space and in time coincides with the relation for the heat equation when we consider the solutions themselves.

The plan of the paper is the following. We begin with a section of notation (section 2). In section 3, we construct a solution of the Navier–Stokes equations as the limit of a sequence of solutions of a linearized approximation of the Navier–Stokes equations. By standard uniqueness results, this solution coincides with the one constructed by Kato in [5]. We recall some essential estimates on the moments of this sequence of approximate solutions and of their derivatives and then we show that these bounds are also valid for the limit solution and its derivatives. The first bounds we obtain are not sufficient for yielding a uniform time decay; they are valid for all time but depend on time. However, owing to the results of [11], we already have uniform bounds for the moments, though not for the moments of the derivatives; for this reason we dedicate section 4 to showing that these moments are also bounded independently of time. The last section deals with the space-time pointwise decay of the solution, which follows from the uniform bound of the moments and an appropriate form of the Gagliardo–Nirenberg inequality.

2. Notation and assumptions.

Let $\alpha = (\alpha_1, \dots, \alpha_n)$ be a multi-index with $\alpha_i \geq 0$. We will use the notation

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}\,, \tag{2.1}$$

where

$$|\alpha| = \alpha_1 + \cdots + \alpha_n\,, \tag{2.2}$$

and

$$D_i = \frac{\partial}{\partial x_i}\,. \tag{2.3}$$

For any integer $m \geq 0$, we set

$$D^m f(x) = \left(\sum_{|\alpha|=m} |D^\alpha f(x)|^2 \right)^{1/2},$$

where $x = (x_1, \ldots, x_n)$. The L^2 norm (or energy norm) will be denoted by

$$\|u\| = \|u(.,t)\|_2 = \left[\int_{\mathbf{R}^n} |u(x,t)|^2 dx\right]^{1/2}, \tag{2.4}$$

where $dx = dx_1 \cdots dx_n$. More generally we denote the L^p norm for $1 \leq p < \infty$ by

$$\|u(.,t)\|_p = \left[\int_{\mathbf{R}^n} |u(x,t)|^p dx\right]^{1/p} \tag{2.5}$$

and the L^∞ norm by

$$\|u(.,t)\|_\infty = \text{ess sup}_x |u(x,t)|. \tag{2.6}$$

The H^m norm is defined by

$$\|u(.,t)\|_{H^m} = \left[\int_{\mathbf{R}^n} \sum_{|\alpha| \leq m} |D^\alpha u(x,t)|^2 dx\right]^{1/2}. \tag{2.7}$$

In what follows, we assume that $u = u(x,t) = (u_1(x,t), \ldots, u_n(x,t))$ is a global solution of the Navier–Stokes equations with the following decay: there exist constants $C, \mu > n/4$ such that

$$\|u(t)\|_2 \leq C(t+1)^{-\mu} \quad \text{for } t \geq 0\,. \tag{2.8}$$

Under these conditions, assuming as always $2 \leq n \leq 5$, it is proved in [12] that the decay given by (2.8) generalizes to

$$\|D^j u(t)\|_2 \leq C(t+1)^{-\mu - j/2} \quad \text{for } t \geq 0\,, j = 0, 1, 2, \ldots\,. \tag{2.9}$$

We recall the Gagliardo–Nirenberg inequality; if $f \in H^m$, then

$$\|D^j f\|_\infty \leq C \|f\|_2^{1-a} \|D^m f\|_2^a\,,$$

with $a = a_{jm} = \frac{j + \frac{n}{2}}{m}$, as long as $j + \frac{n}{2} < m$. Taking m large enough (assuming we can do this) we get from (2.8) and (2.9)

$$\|D^j u(t)\|_\infty \leq C(t+1)^{-\mu - j/2 - n/4} \quad \text{for } j = 0, 1, \ldots\,. \tag{2.10}$$

Combining (2.9) and (2.10) we get as in [12]

$$\|D^j u(t)\|_p \leq C(t+1)^{-\mu - j/2 - n/4(1-2/p)} \quad \text{for } j = 0, 1, \ldots \tag{2.11}$$

for $p \in [2, \infty]$, $t > 0$.

Since we are interested in decay of derivatives and hence in smooth solutions, we are going to work with solutions that start with small data, or the results we establish will only be valid for large t.

The main idea in order to obtain pointwise decay is to prove decay of the moments and then combine this with an appropriate Gagliardo–Nirenberg inequality to yield decay in L^∞, whence the pointwise decay. With this in mind, we introduce the following weighted spaces:

$$f \in L^{r_1}_\nu \quad \text{iff} \quad \left(\int_{\mathbf{R}^n} |x|^{\nu r_1} |f|^{r_1}\, dx\right)^{1/r_1} < \infty. \tag{2.12}$$

For $s = 0, 1, 2, \ldots$, we define the (s, α) moments

$$M_{s,\alpha}(t) = \int_{\mathbf{R}^n} |x|^s |D^\alpha u(x,t)|^2 \, dx,$$

and in particular for $s \geq 0$, $t \geq 0$, we define the moment of order s of u by

$$M_s((u)(t)) = M_{s,0}(t) = \int_{\mathbf{R}^n} |x|^s |u(x,t)|^2 \, dx = \Big(\|u(t)\|_{L^2_{s/2}} \Big)^2 .$$

Finally, define for $s, m = 0, 1, 2, \ldots$,

$$\tilde{M}_{s,m}(t) = \sum_{|\alpha|=m} M_{s,\alpha}(t) = \int_{\mathbf{R}^n} |x|^s |D^m u(x,t)|^2 \, dx.$$

3. Preliminaries. To start our calculations we need to recall some weighted-norms estimates satisfied by approximate solutions to the Navier–Stokes equations [11]. These solutions satisfy a "linearized Navier–Stokes equation," in which both the convective and the pressure terms are linearized in "explicit form." To this purpose, the pressure is expressed as a product of Riesz transforms. Specifically, we construct the sequence $\{u^\ell\}$ of approximate solutions as follows: $v = u^{\ell+1}$ is the solution of

$$\begin{aligned} v_t - \Delta v + u^\ell \cdot \nabla v + \nabla P(u^\ell, v) &= 0, \\ \operatorname{div} v &= 0, \\ v(0) &= u_0 \, , \end{aligned} \tag{3.1}$$

with initial approximation $u^0 = u_0$ and u_0 in an appropriate space. The solution v is constructed locally by a fixed-point argument and then is extended by a priori estimates. It is unique by construction. The bilinear operator P is defined by

$$P(u, v) = \sum_{j,k} R_j R_k (u_j v_k) \, ,$$

where $u = (u_1, \ldots, u_n)$, $v = (v_1, \ldots, v_n)$ are functions from $\mathbf{R}^n$ to $\mathbf{R}^n$, and R_j denotes the Riesz transforms,

$$\widehat{[R_j f]}(\xi) = -i \frac{\xi_j}{|\xi|} \hat{f}(\xi) \quad \text{for } 1 \leq j \leq n \, .$$

When $u^\ell = v$ we recover the Navier–Stokes equations, since the pressure p and the velocity u of the Navier–Stokes equations are related by

$$\Delta p = - \sum_{j,k} \frac{\partial^2}{\partial x_j \partial x_k} (u_j u_k) \, ,$$

hence

$$\hat{p}(\xi, t) = - \sum_{j,k} \frac{\xi_j \xi_k}{|\xi|^2} \widehat{u_j u_k}(\xi, t),$$

and

$$p = \sum_{j,k} R_j R_k (u_j u_k) = P(u, u).$$

The linearization (3.1) is of the type used by Caffarelli, Kohn, and Nirenberg in [1], by Kajikiya and Miyakawa in [4], by Leray in [9], and by Sohr, von Wahl, and Wiegner in [14]. The advantage of making the linearization explicit is that we can apply to the sequence $\{u^\ell\}$ well-known properties of the Riesz transforms, such as their boundedness in L^p-spaces (see [15]) and in weighted L^p-spaces satisfying the Muckenhoupt condition (see [3], [16]), in order to obtain bounds for the solutions of the Navier–Stokes equations and of their moments. We expect that our proofs for establishing bounds in weighted L^p-spaces, with some modifications, could be used for the approximating solutions constructed by Caffarelli, Kohn, and Nirenberg [1], by Kajikiya and Miyakawa [4], and by Sohr, von Wahl, and Wiegner [14].

In [11] we constructed the solution to (3.1) via a fixed-point method. We recall briefly the construction, referring to [11] for details. Let

$$F(x,t) = F(t)(x) = (4\pi t)^{-n/2} e^{-|x|^2/4t}$$

be the fundamental solution of the heat equation in n space variables and set

$$H(u,v) = u \cdot \nabla v + \nabla P(u,v)\,.$$

If v solves (3.1), then v has the expression

$$v(t) = F(t) * u_0 - \int_0^t F(t-s) * H(u^\ell, v)(s)\, ds. \tag{3.2}$$

For $u, \varphi \in L^2([0,T], H^1(\mathbf{R}^n)^n)$, we define

$$\begin{aligned} \mathcal{M}_u\varphi(t) &= \int_0^t F(t-s) * [u\cdot\nabla\varphi(s) + \nabla P(u,\varphi)(s)]\; ds \\ &= \int_0^t F(t-s) * H(u,\varphi)(s)\, ds \end{aligned} \tag{3.3}$$

and

$$\mathcal{L}_u\varphi(t) = F(t) * u_0 - \mathcal{M}_u\varphi(t). \tag{3.4}$$

The integral version (3.2) of (3.1) linearized with respect to u^ℓ becomes

$$v = \mathcal{L}_{u^\ell}(v);$$

that is, the solution to the linearized Navier–Stokes equation (3.1) can be obtained as a fixed point of the operator $\mathcal{L}_{u^\ell}$ (see [11]). We prove in [11] that for some $T > 0$, $T = \infty$ for small data, the sequence $\{u^\ell\}$ converges in $C([0,T], L^2 \cap L^r)$ to a weak solution of the Navier–Stokes equations, provided the data is in $L^2 \cap L^r$ and $r > n$. If the data is also in H^1 and is sufficiently small, the solution will be smooth. These are the solutions we will be interested in. Although Kato [5] has obtained smooth solutions with small data in $L^2 \cap L^n$, we do not use his construction because we want to ensure that the solutions also lie in the appropriate weighted space whenever the data belong to that space too. However, our solutions are clearly Hopf–Leray solutions (see [11, Theorem 2.4]); furthermore, in the notation used by Fabes, Jones, and Riviere in [2], they are in $L^{p,\tilde{q}}(\mathbb{R}^n \times (0,T))$ for every $T > 0$ for some $r > n, \tilde{q} > 2$

(see (3.6) below). By the uniqueness results of section V of [2] our solutions coincide with the solutions of Kato.

In [11] we needed to introduce numbers ν, q, r, r_1 satisfying the relations

$$0 \leq \nu < n, \quad 2 \leq r_1 \leq r, \quad 1 \leq q \leq \infty, \quad r > n\,, \tag{3.5}$$

$$\frac{1}{q} < \frac{\nu}{2} - \frac{n}{2r} + \frac{1}{2}, \quad \frac{1}{r} \leq \frac{1}{r_1} + \frac{\nu}{n} < 1 - \frac{1}{r}\,. \tag{3.6}$$

We recall Lemma 2.2 of [11], which we state here for convenience:

Lemma 3.1. *Assume the function u satisfies*

$$u \in C([0,T], W^{m,r}(\mathbf{R}^n)^n) \cap L^q([0,T], (W^{m,r_1}(\mathbf{R}^n)^n).$$

There exists a constant $K(T,u)$ of the form

$$K(T,u) = C(T)\left(\|u\|_{C_T(W^{m,r})} + \|u\|_{L^q_T(W^{m,r_1})}\right)$$

with $C(T)$ independent of u such that if $D^\alpha u_0 \in L^{r_1}_\nu \cap L^r(\mathbf{R}^n)^n$ for $|\alpha| \leq m$, then the fixed point v of $\mathcal{L}_u$ satisfies $D^\alpha v \in C([0,T], L^{r_1}_\nu(\mathbf{R}^n))$ for $|\alpha| \leq m$ and

$$\|D^\alpha v(t)\|_{L^{r_1}_\nu} \leq C(T)\left(\|u_0\|_{W^{m,r_1}} + \sum_{|\beta|\leq m} \|D^\beta u_0\|_{L^{r_1}_\nu}\right) + K(T,u)\left(\|u_0\|_{W^{m,r}} + \|u_0\|_{W^{m,r_1}}\right).$$

If u is a strong solution of the Navier–Stokes equations, then u is the fixed point of $\mathcal{L}_u$; moreover, $r, r_1 \geq 2$ so that $K(T,u)$ is finite if $2 \leq n \leq 5$ by (2.11) for every $T > 0$. The following corollary is immediate.

Corollary 3.2. *Assume $2 \leq n \leq 5$, conditions (3.5), (3.6), and let u be a strong solution of the Navier–Stokes equations with data $u_0 \in W^{m,r} \cap W^{m,r_1} \cap H^1(\mathbf{R}^n)^n$. Then*

$$\|D^\alpha u(t)\|_{L^{r_1}_\nu} \leq C(T)C_0, \tag{3.7}$$

where C_0 depends only on appropriate norms of the data.

4. Decay of moments of derivatives. In order to obtain the decay of moments of derivatives, we will first need to establish uniform bounds. Once these are obtained, the decay will follow by a Hölder inequality between the (m,s) moments and the L^2 norm of the derivatives.

Theorem 4.1. *Let u_0 be as in Corollary 3.2. Let u be a strong solution of the Navier–Stokes equations with data u_0 satisfying*

$$\|u(t)\|_2 \leq C(t+1)^{-\mu}, \quad \textit{where } \mu > \frac{n}{4} - \frac{1}{2}\,. \tag{4.1}$$

Then

$$\tilde{M}_{s,m}(t) \leq C(t+1)^{-(2\mu+m)(1-\frac{s}{n})}\,, \tag{4.2}$$

for $m = 0, 1, 2, \ldots,$ $s = 0, 1, \ldots, n$.

Proof. As before we note that if the data is sufficiently small then this solution u exists. In particular, if $u \in H^2 \cap L^\infty$, then all the derivatives of higher order are in L^6 (see [12]). Moreover, inequalities (2.9), (2.10), and (2.11) will hold.

For the proof, note that the case $s = 0$ is covered by (2.9). Assuming $s > 0$ from now on, we proceed by induction on m. Each induction step is dealt with following the approach of Theorem 4.1 of [11] (where the case $m = 0$ is proved). As in [11, Theorem 4.1] the estimate for $0 < s < n$ will follow from the estimates for $s = 0$ and $s = n$ by Hölder interpolation. Indeed, let $1/p = (n-s)/n$, $1/p' = s/n$, and $|\alpha| = m$; we have

$$\begin{aligned} M_{s,\alpha}(t) = \int_{\mathbf{R}^n} |x|^s |D^\alpha u|^2\, dx &\le \left(\int_{\mathbf{R}^n} |D^\alpha u|^2\, dx\right)^{1/p} \left(\int_{\mathbf{R}^n} |x|^n |D^\alpha u|^2\, dx\right)^{1/p'} \\ &= M_{0,\alpha}(t)^{1-\frac{s}{n}} M_{n,\alpha}(u)(t)^{\frac{s}{n}} \le C(t+1)^{-(2\mu+m)(1-\frac{s}{n})} M_{n,\alpha}(u)(t)^{\frac{s}{n}}. \end{aligned}$$

Thus, if $M_{n,\alpha}(u)(t)$ is uniformly bounded, we have

$$\tilde{M}_{s,m}(t) \le C(t+1)^{-(2\mu+m)(1-\frac{s}{n})}\,.$$

It suffices thus to prove the estimate for $s = n$, which merely says that $\tilde{M}_{n,m}(t)$ is bounded uniformly with respect to t, for $t > 0$. In other words, it suffices to prove

$$\sup_{t>0} \tilde{M}_{n,m}(t) < \infty \tag{4.3}$$

for $m = 0, 1, \dots$.

Let α be a multi-index with $|\alpha| = m$. For a function g and a multi-index β, we set $g_\beta = D^\beta g$. By Leibniz's product formula, differentiating (1), we obtain

$$u_{\alpha t} = \Delta u_\alpha - \sum_{\beta+\gamma=\alpha} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} u_\beta \cdot \nabla u_\gamma - \nabla p_\alpha;$$

dot multiplying by $|x|^s u_\alpha$ and using that $\operatorname{div} u = 0$ and $\operatorname{div} u_\alpha = 0$, we get, after some technical but straightforward manipulations,

$$\begin{aligned} |x|^s u_{\alpha t} \cdot u_\alpha = &-|x|^s |\nabla u_\alpha|^2 + \frac{s}{2}(s-2+n)|x|^{s-2}|u_\alpha|^2 + \frac{s}{2}|x|^{s-2}(x\cdot u)|u_\alpha|^2 \\ &-|x|^s \sum_{\beta+\gamma=\alpha,\beta\neq 0} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} (u_\beta \cdot \nabla u_\gamma)\cdot u_\alpha + s|x|^{s-2}(x\cdot u_\alpha) p_\alpha \\ &+ \operatorname{div} E_{s,\alpha}\,, \end{aligned}$$

where

$$E_{s,\alpha} = \frac{|x|^s}{2}\nabla(|u_\alpha|^2) - \frac{s}{2}|x|^{s-2}|u_\alpha|^2 x - \frac{|x|^s}{2}|u_\alpha|^2 u - |x|^s u_\alpha p_\alpha.$$

One can prove now, as in Lemma 6.1, Appendix B of [11], that

$$\liminf_{R\to\infty} \int_{|x|=R} |E_{s,\alpha}|\, dS = 0.$$

More precisely, the proof is a repetition of the arguments in the above mentioned lemma, where we replace u by u_α and use the appropriate estimates for the derivatives obtained in [12]. Thus

$$\int_{\mathbf{R}^n} \operatorname{div} E_{s,\alpha}\, dx = 0\,,$$

and we obtain

$$\frac{1}{2}\frac{d}{dt}M_{s,\alpha}(t) = A(t) + B(t) + C(t) + D(t), \tag{4.4}$$

where

$$\begin{aligned} A(t) &= -\int_{\mathbf{R}^n} |x|^s |\nabla u_\alpha|^2\,dx + \frac{s}{2}(s-2+n)M_{s-2,\alpha}(t)\frac{s}{2}(s-2+n)M_{s-2,\alpha}(t),\\ B(t) &= \frac{s}{2}\int_{\mathbf{R}^n} |x|^{s-2}(x\cdot u)|u_\alpha|^2\,dx,\\ C(t) &= -\sum_{\beta+\gamma=\alpha,\beta\neq 0}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}\int_{\mathbf{R}^n}|x|^s(u_\beta\cdot\nabla u_\gamma)\cdot u_\alpha\,dx,\\ D(t) &= s\int_{\mathbf{R}^n}|x|^{s-2}(x\cdot u_\alpha)p_\alpha\,dx. \end{aligned}$$

Assume $m = 0$. Recall that we write M_s for $M_{s,0}$. We prove by induction on s that there exists $C \geq 0$ such that $M_s(u)(t) \leq C$ for all $t \geq 0$, $s = 1, \ldots, n$. We begin considering the case $s = 2$; the case $s = 1$ follows by interpolation between the cases $s = 0$ and $s = 2$ and induction can then proceed in steps of 2; i.e., M_k bounded implies M_{k+2} bounded.

If A, B, C, D are as in (4.4) for $|\alpha| = m = 0$, $s = 2$, we get $A(t) \leq nM_{0,0}(u)(t) = n\|u(t)\|_2^2$, $C(t) = 0$ and

$$\begin{aligned} B(t) &\leq \int_{\mathbf{R}^n}|x||u|^3\,dx \leq M_2(u)(t)^{1/2}\|u(t)\|_4^2,\\ D(t) &\leq M_2(u)(t)^{1/2}\|p(t)\|_2^2 \leq CM_2(u)(t)^{1/2}\|u(t)\|_4^2 \end{aligned}$$

so that by (4.4)

$$\frac{d}{dt}M_2(u)(t) \leq CM_2(u)(t)^{1/2}\|u(t)\|_4^2 + n\|u(t)\|_2^2. \tag{4.5}$$

By (2.11) (with $j = 0$ and $p = 4$)

$$\|u(t)\|_4 \leq C(t+1)^{-\mu-n/8};$$

it follows from this and (4.1)

$$\frac{d}{dt}M_2(u)(t) \leq n(t+1)^{-2\mu} + CM_2(u)(t)^{1/2}(t+1)^{-\delta}$$

with $\delta = 2\mu + n/4$. A bit of elementary arithmetic yields

$$\frac{d}{dt}M_2(u)(t) \leq n(t+1)^{-2\mu} + C(t+1)^{-\delta} + CM_2(u)(t)(t+1)^{-\delta}.$$

Since the moments are bounded (time dependent) and since $\delta > 1$ it follows by Gronwall's inequality that

$$M_2(u)(t) \leq Ce^{\int_0^\infty (t+1)^{-\delta}dt} \leq \text{const}$$

proving the case $s = 2$. Assume now $s > 2$. In this case

$$A(t) \leq \frac{s}{2}(s + n - 2)M_{s-2}(u)(t), \tag{4.6}$$

$$B(t) \leq \int_{\mathbf{R}^n} |x|^{s-1}|u|^3\,dx \leq M_s(u)(t)^{(s-1)/s}\|u(t)\|_{s+2}^{(s+2)/s}, \tag{4.7}$$

$$D(t) \leq \int_{\mathbf{R}^n} |x|^{s-1}|u||p|\,dx \leq CM_s(u)(t)^{(s-1)/s}\|u(t)\|_{s+2}^{(s+2)/s}. \tag{4.8}$$

Inequality (4.7) is an immediate consequence of Hölder's inequality (with exponents $s/(s-1)$ and s). For (4.8) notice first that, by Hölder's inequality,

$$\begin{aligned}\int_{\mathbf{R}^n} |x|^{s-1}|u||p|\,dx &\leq \left(\int_{\mathbf{R}^n} |x|^s|u|^2\,dx\right)^{1/2}\left(\int_{\mathbf{R}^n} |x|^{s-2}|p|^2\,dx\right)^{1/2}\\ &= M_s(u)(t)^{1/2}\|p\|_{L^2_\nu}\end{aligned}$$

with $\nu = s/2 - 1$. Then $n/(n-\nu) < 2$ (since $s < n+2$) hence the Riesz transforms are bounded in L^2_ν (see [11, Lemma 5.1]); since $p = -\sum_{j,k} R_jR_k(u_ju_k)$, we get

$$\begin{aligned}\|p\|_{L^2_\nu} &\leq C\|\,|u|^2\|_{L^2_\nu}\\ &= C\left(\int_{\mathbf{R}^n} |x|^{s-2}|u|^4\,dx\right)^{1/2} \leq M_s(u)(t)^{(s-2)/2s}\|u\|_{s+2}^{(s+2)/s},\end{aligned}$$

where we factored $|u|^4 = |u|^{2-4/s}|u|^{2+4/s}$ and used Hölder's inequality with exponents $s/(s-2)$, $s/2$. Inequality (4.8) follows. To continue estimating, we get by Hölder's inequality, (4.1), and (2.10) (for $j = 0$)

$$\|u(t)\|_{s+2} \leq \|u(t)\|_2^{2/(s+2)}\|u\|_\infty^{s/(s+2)} \leq C(t+1)^{-\mu-(ns)/(4s+8)}\,;$$

by Hölder's inequality and (4.1)

$$M_{s-2}(u)(t) \leq M_s(u)(t)^{(s-2)/s}\|u(t)\|_2^{4/s} \leq C(t+1)^{-4\mu/s}M_s(u)(t)^{(s-2)/s};$$

whence combining with (4.6), (4.7), (4.8), and (4.4),

$$\begin{aligned}\frac{d}{dt}M_s(u)(t) &\leq C_1(t+1)^{-4\mu/s}M_s(u)(t)^{(s-2)/s}\\ &\qquad + C_2(t+1)^{-\mu(s+2)/s-n/4}M_s(u)(t)^{(s-1)/s}.\end{aligned}$$

We estimate the two terms on the right-hand side (R.H.S.) using $M_s^\tau \leq 1 + M_s$ for $\tau = (s-2)/s$ and $\tau = (s-1)/s$, respectively; we get

$$\frac{d}{dt}M_s(u)(t) \leq C_1(t+1)^{-\rho} + C_2(t+1)^{-\rho}M_s(u)(t),$$

where

$$\rho = \min\left\{\frac{4\mu}{s}, \frac{s+2}{s}\mu + \frac{n}{4}\right\} > 1.$$

Integrating from 0 to t, considering that

$$\int_0^t (\sigma+1)^{-\rho}\,d\sigma \leq \int_0^\infty (\sigma+1)^{-\rho}\,d\sigma = \frac{1}{\rho - 1} < \infty,$$

we get

$$M_s(u)(t) \leq C\,(1 + M_s(u)(0)) + C\int_0^t (\sigma+1)^{-\rho} M_s(u)(\sigma)\,d\sigma.$$

By Gronwall's lemma,

$$M_s(u)(t) \leq C\,(1 + M_s(u)(0))\, e^{\int_0^\infty (\sigma+1)^{-\rho}\,d\sigma} \leq C_0 < \infty$$

which completes the proof of the case $m = 0$.

Assume now m is a positive integer and that the estimates (4.2) have been proved up to $m-1$; $s = 0, \ldots, n$. Let $|\alpha| = m$. Time dependent bounds for $M_{s,\alpha}(u)(t)$ are easily established by induction on s, $0 \leq s \leq n$. In fact, the case $s = 0$ is (as already mentioned) immediate and the induction proceeds by means of energy estimates which are quite straightforward and as such will be omitted; the reader can refer to [11] for details of a similar proof. With this established, to obtain the uniform bound we proceed as follows. Let $A(t)$, $B(t)$, $C(t)$, $D(t)$ be as in (4.4) with $s = n$.

Bound for $A(t)$.

Notice first that if $n = 2$ then

$$A(t) \leq 2M_{0,\alpha}(t) \leq C_0(1+t)^{-2\mu-1}\,, \tag{4.9}$$

where $2\mu > n/2 = 1$. Suppose now that $3 \leq n \leq 5$; by Hölder's inequality and by (2.9),

$$M_{n-2,\alpha}(t) \leq M_{n,\alpha}(t)^{(n-2)/n}\|u_\alpha(t)\|_2^{4/n} \leq C(1+t)^{-\rho} M_{n,\alpha}(t)^{(n-2)/n},$$

with $\rho = (4/n)(\mu + m/2) > 1$. In general, from now on, ρ denotes a constant > 1, not the same one in all inequalities. By the definition of $A(t)$, using also

$$(1+t)^{-\rho} M_{n,\alpha}(t)^{(n-2)/n} \leq \frac{2}{n}(1+t)^{-\rho} + \frac{n-2}{n}(1+t)^{-\rho} M_{n,\alpha}(t),$$

we prove that

$$|A(t)| \leq C(1+t)^{-\rho}\left(1 + \tilde{M}_{n,m}(t)\right)\,. \tag{4.10}$$

Bound for $B(t)$.

$$\begin{aligned} |B(t)| &= \left|\frac{n}{2}\int_{\mathbf{R}^n} |x|^{n-2}(x\cdot u)|u_\alpha|^2\,dx\right| \leq \frac{n}{2}\int_{\mathbf{R}^n} |x|^{n-1}|u||u_\alpha|^2\,dx \\ &\leq \frac{n}{2}\|u_\alpha\|_2^{2/n}\|u\|_\infty \left(\tilde{M}_{n,m}(t)\right)^{(n-1)/n}\,, \end{aligned}$$

so that by (2.9) and (2.10),

$$\begin{aligned} |B(t)| &\leq C(1+t)^{-\rho}\left(\tilde{M}_{n,m}(t)\right)^{(n-1)/n} \\ &\leq C(1+t)^{-\rho}\left(1 + \tilde{M}_{n,m}(t)\right), \end{aligned}$$

where this time $\rho = (2/n)(\mu + m/2) + \mu + n/4 > 1$.

Bound for $C(t)$.

Note that C(t) is a sum in terms of α and β, where $|\beta| + |\gamma| = |\alpha|$ and $\beta \neq 0$. The general term in $C(t)$ can be estimated by

$$\int_{\mathbf{R}^n} |x|^n \, |(u_\beta \cdot \nabla u_\gamma) \cdot u_\alpha| \, dx \leq \|D^j u\|_\infty \tilde{M}_{n,\ell}(t)^{1/2} \tilde{M}_{n,m}(t)^{1/2},$$

where $j = \min(|\beta|, |\gamma| + 1)$, $\ell = \max(|\beta|, |\gamma| + 1)$, so that $0 \leq j \leq [m/2]$, $[(m+1)/2] \leq \ell \leq m$, and $j + \ell = m + 1$. When $\ell = m$, and so $j = 1$, (2.10) implies a bound of the form

$$C(1+t)^{-(\mu+n/4+1)} \tilde{M}_{n,m}(t).$$

The terms with $\ell < m$ are bounded, using the induction hypothesis and (2.10), by

$$C(1+t)^{-(\mu+n/4+j/2)} \tilde{M}_{n,m}(t)^{1/2},$$

and we obtain again an estimate of the form

$$|C(t)| \leq C(1+t)^{-\rho} \left(1 + \tilde{M}_{n,m}(t)\right), \tag{4.11}$$

where $\rho > 1$.

Bound for $D(t)$.

Since the Riesz transforms are bounded in L^2_ν with $\nu = (n-2)/2$, and D^α commutes with the Riesz transforms, we can write

$$p_\alpha = D^\alpha p = \sum_{j,k} R_j R_k [D^\alpha (u_j u_k)] = \sum_{k,j,\beta+\gamma=\alpha} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} R_j R_k (u_{\beta,j} u_{\gamma,k}),$$

and we have

$$\begin{aligned} |D(t)| &= \left| n \int_{\mathbf{R}^n} |x|^{n-2} (x \cdot u_\alpha) p_\alpha \, dx \right| \\ &\leq C \int_{\mathbf{R}^n} |x|^{n-1} |u_\alpha| |p_\alpha| \, dx \leq C \tilde{M}_{n,m}(t)^{1/2} \|p_\alpha\|_{L^2_\nu} \\ &\leq C \tilde{M}_{n,m}(t)^{1/2} \sum_{\beta+\gamma=\alpha} \| |u_\beta| |u_\gamma| \|_{L^2_\nu} . \end{aligned}$$

Then Hölder's inequality gives

$$\begin{aligned} \| |u_\beta| |u_\gamma| \|_{L^2_\nu} &= \left(\int_{\mathbf{R}^n} |x|^{n-2} |D^\beta u|^2 |D^\gamma u|^2 \, dx \right)^{1/2} \\ &\leq C \|D^j u\|_\infty \|D^\ell u\|_2^{1/n} \tilde{M}_{n,\ell}(t)^{(n-2)/2n}, \end{aligned}$$

with $j = \min(\beta, \gamma)$, $\ell = \max(\beta, \gamma)$ (so $0 \leq j \leq m/2$). Once more we apply (2.9), (2.10) to get $\|D^j u\|_\infty \|D^\ell u\|_2^{1/n} \leq C(1+t)^{-\rho}$ with $\rho = (1/n)(\mu + \ell/2) + \mu + n/4 + j/2 > 1$. By the induction hypothesis $\tilde{M}_{n,\ell}(t)$ is bounded uniformly in t if $\ell < m$, so all terms with $\ell < m$ in the last estimate for D can be bounded by $C(1+t)^{-\rho}$ and the remaining term is bounded by

$$C(1+t)^{-\rho} \tilde{M}_{n,m}(t)^{(n-2)/2n} \leq C(1+t)^{-\rho} \left(1 + \tilde{M}_{n,m}(t)\right)$$

so that $D(t)$ has a bound of the same type as $A(t)$, $B(t)$, $C(t)$. Combining all the above estimates, we derive

$$\frac{d}{dt}\tilde{M}_{n,m}(t) \leq C(1+t)^{-\rho} + C(1+t)^{-\rho}\tilde{M}_{n,m}(t)\,,$$

where $\rho > 1$. Hence, integrating in this inequality, we find

$$\tilde{M}_{n,m}(t) \leq \left(\tilde{M}_{n,m}(0) + \frac{C}{\rho-1}\right) + C\int_0^t (s+1)^{-\rho}\tilde{M}_{n,m}(s)\,ds.$$

Then Gronwall's lemma implies

$$\tilde{M}_{n,m}(t) \leq \left(\tilde{M}_{n,m}(0) + \frac{C}{\rho-1}\right) e^{c/(\rho-1)}\,,$$

thus proving that $\tilde{M}_{n,m}(t)$ is bounded uniformly with respect to t for $t > 0$. □

Note. We took some pains to avoid having to bound $\|D^j u\|_\infty$ for $j > [(m+1)/2]$. In this way, bounds on the L^2-norm of derivatives of order m will give (sometimes) all the needed L^∞ bounds on the $D^j u$'s.

The next theorem establishes the spatial and time decay of strong solutions to equations for which the moments decay.

Theorem 4.2. *Let $2 \leq n \leq 5$. With the assumptions of Theorem 4.1, let u be a strong solution u of the Navier–Stokes equations with data u_0. Let $k \leq n/2$. Then*

$$|D^\alpha u(x,t)| \leq C_{k,m}\frac{1}{(t+1)^{\rho_0}(1+|x|^2)^{k/2}}\,, \tag{4.12}$$

where $\rho_0 = (\mu + m/2 + n/4)(1-2k/n)$ and $|\alpha| = m$.

Proof. Note that n is restricted to the values $2 \leq n \leq 5$ for which we have estimates for the moments. The main tools for the proof are Theorem 4.1 and the Gagliardo–Nirenberg inequality. Let

$$v(x,t) = (1+|x|^2)^{k/2}D^\alpha u(x,t)\,.$$

By Leibniz's formula, we have

$$D^s v = \sum_{j=0}^{s} c_j^s (1+|x|^2)^{\frac{k-j}{2}} D^{s-j}u_\alpha\,. \tag{4.13}$$

Together with the decay of the moments of derivatives given by Theorem 4.1, this formula implies that

$$\|D^s v\|_2 \leq C_0\sum_{j=0}^{s}(1+t)^{-(\mu+m/2+(s-j)/2)(1-2(k-j)/n)}\,. \tag{4.14}$$

Since the function $f(j) = (\mu + m/2 + (s-j)/2)(1-2(k-j)/n)$ is increasing, it has a minimum at $j = 0$. Thus we have

$$\|D^s v\|_2 \leq C_0(1+t)^{-(\mu+m/2+s/2)(1-2k/n)}\,. \tag{4.15}$$

In particular when $s = 0$,

$$\|v\|_2 \leq C_0(1+t)^{-(\mu+m/2)(1-2k/n)}\,. \tag{4.16}$$

Let us apply the Gagliardo–Nirenberg inequality with $a = n/(2s) < 1$, provided $n/2 < s$, i.e., $s > [n/2]$, to get

$$\|v(\cdot,t)\|_\infty \leq \|v(\cdot,t)\|_2^{1-a} \|D^s v(\cdot,t)\|_2^a \,. \tag{4.17}$$

Combining with (4.16) and (4.15) yields

$$|(1+|x|^2)^{k/2} D^\alpha u(x,t)| \leq ||v(t)||_\infty \leq C_0 (1+t)^{-\rho_0} \,,$$

where

$$\begin{aligned} \rho_0 &= (1-2k/n)\left((\mu + m/2 + s/2)n/(2s) + (1-n/(2s))(\mu + m/2)\right) \\ &= (1-2k/n)(\mu + m/2 + n/4) \,. \end{aligned}$$

We note that the above value of ρ_0 is independent of s. Thus we could have obtained it using only the s derivative with $s > [n/2]$. In particular note that when $n = 3$, it suffices to use $s = 2$ and $\rho_0 = (\mu + m/2 + 3/4)(1 - 2k/3)$. The proof is complete. □

4.1. Comparison with the heat equation. It is easy to show that the fundamental solution of the heat equation,

$$E(x,t) = (4\pi t)^{-n/2} e^{-|x|^2/4t} \,,$$

which is the linear part of the Navier–Stokes equations, has the following asymptotic behavior:

$$|D^\alpha E(x,t)| \leq c_0 |x|^{-a} t^{-b},$$

where $a + 2b = n + m$, with $m = |\alpha|$. It is also easy to show that there is a large class of solutions to the heat equation which will have the same type of decay. For instance solutions such that the data satisfies $u_0 \in \mathcal{K}$ where

$$\mathcal{K} = \{u_0 : u_0(y) \geq e^{-y^2/4t_0}\}$$

will have the above type of decay, provided we are considering $t \geq t_0 + \varepsilon$. In the case of solutions to the Navier–Stokes equations, if we take $\mu = n/4$, the relation that holds between the decay in space and in time is

$$2\rho_0 + 2k = m + n - \frac{2km}{n} \,.$$

For $k = 0$, we recover the decay of the heat equation, but this only gives decay in time. If $m = 0$ we recover the relation $2\rho_0 + 2k = n$; i.e., we have the same decay relation in space and in time as for solutions to the heat equation.

Final remarks. We expect that our results can be extended easily to dimensions 6 and 7 using the L^2 decay results, for derivatives of higher order, recently obtained by Wiegner [19].

Acknowledgment. The authors wish to thank the referee for many helpful suggestions which resulted in a significantly improved paper.

REFERENCES

[1] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.

[2] E. Fabes, B. Jones, and N. Riviere, *The initial value problem for the Navier-Stokes equations with data in L^p*, Arch. Rational Mech. Anal., 45 (1972), pp. 222–240.

[3] J.-L. Journé, *Calderón-Zygmund Operators, Pseudo-Differential Operators and the Cauchy Integral of Calderón*, Lecture Notes in Math. 994, Springer-Verlag, Berlin, Heidelberg, 1983.

[4] R. Kajikiya and T. Miyakawa, *On the L^2 decay of weak solutions of the Navier-Stokes equations in $\mathbf{R}^n$*, Math. Z., 192 (1986), pp. 135–148.

[5] T. Kato, *Strong L^p solutions of the Navier-Stokes equations with applications to weak solutions*, Math. Z., 187 (1982), pp. 471–480.

[6] G. H. Knightly, *On a class of global solutions of the Navier-Stokes equations*, Arch. Rational Mech. Anal., 21 (1966), pp. 211–245.

[7] H. Kozono, *Global L^n solutions and its decay property for the Navier-Stokes equations in half space $\mathbf{R}^n$*, J. Differential Equations, 79 (1989), pp. 79–88.

[8] H. Kozono and T. Ogawa, *Two dimensional Navier-Stokes equations in unbounded domains*, Math. Ann., 297 (1993), pp. 1–31.

[9] J. Leray, *Essai sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.

[10] M. Schonbek, *Lower bounds of rates of decay for solutions to the Navier-Stokes equations*, J. Amer. Math. Soc., 4 (1991), pp. 423–449.

[11] M. Schonbek and T. Schonbek, *On the Boundedness and Decay of Moments of Solutions of the Navier-Stokes Equations*. preprint, 1998.

[12] M. Schonbek and M. Wiegner, *On the decay of higher order norms of the solutions of Navier-Stokes equations*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 677–685.

[13] M. E. Schonbek, *Large time behaviour of solutions to the Navier-Stokes equations*, Comm. Partial Differential Equations, 11 (1986), pp. 733–763.

[14] H. Sohr, W. von Wahl, and M. Wiegner, *Zur Asymptotik der Gleichungen von Navier-Stokes*, Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II, 3 (1986), pp. 45–59.

[15] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[16] E. Stein, *Harmonic Analysis: Real Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[17] S. Takahashi, *A Weighted Equation Approach to Decay Rate Estimates for the Navier-Stokes Equations*. preprint, 1997.

[18] M. Wiegner, *Decay results for weak solutions to the Navier-Stokes equations in $\mathbf{R}^n$*, J. London Math. Soc. (2), 35 (1987), pp. 303–313.

[19] M. Wiegner, *Higher order estimates in further dimensions for the solutions to the Navier-Stokes equations*, in Proceedings of the Banach Center, 1999.

[20] L. Zhang, *Sharp Rates of Decay of Global Solutions to* 2*-Dimensional Navier-Stokes Equations*. preprint, 1994.

SIAM J. Math. Anal.
Vol. 31, No. 4, pp. 754–775

ASYMPTOTIC EXPANSIONS OF SYMMETRIC STANDARD ELLIPTIC INTEGRALS[*]

JOSÉ L. LÓPEZ[†]

Abstract. Symmetric standard elliptic integrals are considered when one of their parameters is larger than the others. The distributional approach is used for deriving five convergent expansions of these integrals in inverse powers of the respective five possible asymptotic parameters. Four of these expansions also involve a logarithmic term in the asymptotic variable. Coefficients of these expansions are obtained by recurrence. For the first four expansions these coefficients are expressed in terms of elementary functions, whereas coefficients of the fifth expansion involve nonelementary functions. The convergence speed of any of these expansions increases for increasing difference between the asymptotic variable and the remaining ones. All the expansions are accompanied by an error bound at any order of the approximation.

Key words. elliptic integrals, asymptotic expansions, distributional approach

AMS subject classifications. 41A60, 33E05

PII. S0036141099351176

1. Introduction. Elliptic integrals (EI) are integrals of the type $\int R(x,y)dx$, where $R(x,y)$ is a rational function of x and y, with y^2 a polynomial of the third or fourth degree in x. When the polynomial y^2 does not have a repeated factor and $R(x,y)$ contains some odd power of y, EI cannot, in general, be expressed in terms of elementary functions. Legendre showed that all EI can be expressed in terms of three standard EI (Legendre's normal EI) [14].

The three complete EI of the first, second, and third kind are particularly important cases of the respective three standard EI. These integrals and the three standard EI are special nonelementary functions that play an important role in several mathematical problems. The first complete EI appears as a certain limit in the theory of iterated number sequences based on the arithmetic geometric mean [18, sect. 12.1.2]. Standard EI are related to theta functions and the Weierstrass elliptic function [18, sect. 12.3]. EI constitute a basic ingredient of certain geometrical [11] and statistical [17] problems.

EI are also involved in several physical problems. The period of a simple pendulum in a constant gravitational field can be expressed in terms of the first complete EI [18, sect. 12.1.1]. The zeros of EI can be used for determining an upper bound for the number of limit cycles of certain Hamiltonian systems [19]. EI are related to certain problems of electromagnetism [20].

A survey of properties of the standard EI can be found, for example, in [1, chap. 17], [2], or [18, chap. 12]. However, as it has been shown by Carlson [5, 6, 7, 8, 9], for numerical computations it is more convenient to use symmetric standard EI instead of Legendre's normal EI. (Legendre's normal EI are connected with the symmetric standard EI by means of simple formulas [18, eq. (12.33)].) A very complete table of the three symmetric standard EI can be found in [5, 6, 7, 8, 9]. They are defined as

[*] Received by the editors February 5, 1999; accepted for publication (in revised form) July 28, 1999; published electronically March 21, 2000. This work was supported by Comisión Interministerial de Ciencia y Tecnología.

http://www.siam.org/journals/sima/31-4/35117.html

[†] Departamento de Matemática e Informatica, Universidad Publica de Navarra, Campus de Arrosadia s/n, 31006 Pamplona, Spain (jl.lopez@unavarra.es).

follows:

$$R_F(x,y,z)=\frac{1}{2}\int_0^\infty \frac{dt}{\sqrt{(t+x)(t+y)(t+z)}},$$

$$R_D(x,y,z)=\frac{3}{2}\int_0^\infty \frac{dt}{\sqrt{(t+x)(t+y)(t+z)^3}},$$

$$R_J(x,y,z,p)=\frac{3}{2}\int_0^\infty \frac{dt}{\sqrt{(t+x)(t+y)(t+z)}(t+p)},$$

where we assume that the parameters x, y, z are nonnegative. We assume also that they are distinct (otherwise these integrals reduce to elementary functions). If the fourth argument of R_J is negative, the Cauchy principal value of R_J can be written in terms of R_F and R_J with all the arguments nonnegative [10]. Therefore, we will consider $p>0$ and $p\neq x,\ y,\ z$ (otherwise R_J reduces to R_D).

On the other hand, the asymptotic approximation of EI has not been exhaustively investigated: classical methods for approximation of integrals cannot be applied. Some results concerning approximations of EI can be found for example in [2] and [13]. However, the more recent results about the asymptotic behavior of these integrals have been obtained by Carlson, Gustafson, and Wong: R_F, R_D, and R_J may be written as a convolution and the method of regularization [22, chap. 6, sect. 7] can be applied.

When one of the parameters of the integrals tends to zero or infinity, the first (and sometimes the second too) term of the asymptotic expansion of R_F, R_D, and R_J, as well as a quite accurate bound for the first error term, has been obtained by Gustafson [12]. Higher terms of the expansion and higher error bounds are not explicitly derived in that work because of the complexity of the Mellin transforms involved in their calculation. Using a very clever analytical trick [10], Carlson and Gustafson have sharpened the bounds for the first error terms obtained in [12] in the case of one parameter going to infinity. Besides, they supply in [10] very accurate bounds for the first error term of the totally symmetric EI of the second kind. Moreover, for all the symmetric EI, they consider also the case of several parameters going to infinity.

Complete convergent expansions of R_F, R_D, and R_J (and not only first terms) have been obtained by Carlson using Mellin transform techniques [3]. Although these expansions have an attractively simple structure, explicit computation of the terms of the expansions is not straightforward and the upper bound on the truncation error is not quite satisfactory [3, sect. 5]. Carlson and Gustafson have solved this problem for $R_F(x,y,z)$ in [4], where an algorithm for computing the coefficients of the convergent expansion of $R_F(x,y,z)$ in terms of Legendre functions and their derivatives is derived. Moreover, accurate error bounds are given at any order of the approximation.

In this paper we try to solve for R_D and R_J the problem that Carlson and Gustafson have solved for R_F. That is, we consider complete convergent expansions for R_D and R_J when one of their parameters x, y, z, or p is large. Then, we face the challenge of obtaining easy algorithms for computing the coefficients of these expansions (in terms of elementary functions when it is possible) and simple expressions for the error bounds at any order of the approximation. For completeness, we also include R_F in this project.

For this purpose, in section 2, we make a review of the asymptotic expansions of Stieltjes [22, chap. 6, sect. 2] and generalized Stieltjes transforms (see [21, Theorem

2 and Example 1]): the distributional approach is used in Lemmas 2.5 and 2.6 and Theorems 2.4 and 2.7 for deriving complete expansions of a certain family of integrals which contains R_F, R_D, and R_J.

On the other hand, using Lemmas 2.8 and 2.9, we obtain simple expressions for the error bounds in the expansions of this family of integrals in Propositions 2.10 and 2.11. In section 3 we apply the results of section 2 for deriving complete convergent expansions of $R_F(x,y,z)$, $R_D(x,y,z)$, $R_D(x,z,y)$, $R_J(x,y,z,p)$, and $R_J(x,y,p,z)$ for large z. They are presented in Corollaries 3.1–3.8 accompanied by error bounds at any order of the approximation. Numerical examples are shown as an illustration. A brief summary and a few comments are postponed to section 4.

2. Distributional approach.

The procedure for deriving convergent expansions of the integrals R_F, R_D, and R_J is based on the distributional approach. It requires the concepts of rapidly decreasing functions and tempered distributions.

DEFINITION 2.1. *We denote by $\mathcal{S}$ the space of rapidly decreasing functions (infinitely differentiable functions $\varphi(t)$ defined on $[0,\infty)$ that, together with their derivatives, approach zero more rapidly than any power of t^{-1} as $t\to\infty$).*

DEFINITION 2.2. *We denote by $\langle\Lambda,\varphi\rangle$ the image of a tempered distribution Λ (a continuous linear functional defined over $\mathcal{S}$) acting over a function $\varphi\in\mathcal{S}$. Recall that we can associate to any locally integrable function $g(t)$ on $[0,\infty)$ a tempered distribution Λ_g defined by*

$$\langle\Lambda_g,\varphi\rangle=\int_0^\infty g(t)\varphi(t)dt.$$

DEFINITION 2.3. *For a locally integrable function $f(t)$ on $(0,\infty)$, we denote by $M[f;w]$ the Mellin transform of $f(t)$ or its analytic continuation. It is defined by*

$$M[f;w]=\int_0^\infty t^{w-1}f(t)dt \tag{2.1}$$

when the integral converges.

The derivation of convergent expansions of $R_J(x,y,z,p)$ for large p is based on the following theorem proved in [22, chap. 6, Theorem 1].

THEOREM 2.4. *Let $f(t)$ be a locally integrable function on $[0,\infty)$, and $\{A_k\}$ be a sequence of complex numbers and let $f(t)$ satisfy, for $n=1,2,3,\dots$,*

$$f(t)=\sum_{k=0}^{n-1}\frac{A_k}{t^{k+\alpha}}+f_n(t),$$

where $f_n(t)=\mathcal{O}(t^{-n-\alpha})$ as $t\to\infty$ and $0<\alpha<1$. Then, for $p>0$ and $n=1,2,3,\dots$,

$$\int_0^\infty\frac{f(t)}{t+p}dt=\frac{\pi}{\sin(\alpha\pi)}\sum_{k=0}^{n-1}(-1)^k\frac{A_k}{p^{k+\alpha}}+\sum_{k=0}^{n-1}(-1)^k\frac{M[f;k+1]}{p^{k+1}}+R_n(p). \tag{2.2}$$

The remainder term satisfies

$$R_n(p)=n!\int_0^\infty\frac{f_{n,n}(t)dt}{(t+p)^{n+1}}, \tag{2.3}$$

where $f_{n,n}(t)$ is defined by

$$f_{n,n}(t)=\frac{(-1)^n}{(n-1)!}\int_t^\infty (u-t)^{n-1}f_n(u)du. \tag{2.4}$$

Convergent expansions of $R_F(x,y,z)$, $R_D(x,y,z)$, $R_D(x,z,y)$, and $R_J(x,y,z,p)$ for large z can be derived from [21, Theorem 2] (see also Example 1 there). This result has been proved by using Mellin transform techniques and, as it is suggested by Wong [21, Example 1], it can also be proved by using the distributional approach. We carry out Wong's proposal in the following two lemmas and Theorem 2.7. The first lemma is proved in [22, chap. 6, Lemma 2].

LEMMA 2.5. *Let $f(t)$ be as in Theorem* 2.4 *but with $\alpha=1$. Then, for any integer $n\geq 1$ and for any function $\varphi\in\mathcal{S}$ we have*

$$\langle f,\varphi\rangle=-\sum_{k=0}^{n-1}\frac{A_k}{k!}\langle\log(t),\varphi^{(k+1)}\rangle+\sum_{k=0}^{n-1}\frac{B_k}{k!}\langle\delta,\varphi^{(k)}\rangle+(-1)^n\langle f_{n,n},\varphi^{(n)}\rangle,$$

where f, $f_{n,n}$, and $\log(t)$ *denote the tempered distributions associated with the locally integrable functions $f(t)$, $f_{n,n}(t)$, and* $\log(t)$*, respectively; δ is the delta distribution in the origin; and*

$$\begin{aligned}B_k&=A_k\sum_{j=1}^{k}\frac{1}{j}+\lim_{w\to k+1}\left\{M[f;w]+\frac{A_k}{w-k-1}\right\}\\&=A_k\sum_{j=1}^{k}\frac{1}{j}+\int_0^1 t^k f_k(t)dt+\int_1^\infty t^k f_{k+1}(t)dt,\end{aligned}\tag{2.5}$$

empty sums being understood as zero.

LEMMA 2.6. *Let $f(t)$ be as in Theorem* 2.4 *with* $0<\alpha\leq 1$. *Define, for* $t\in[0,\infty)$, $z>0$, $\eta>0$, *and* $\alpha+\rho>1$,

$$\varphi_\eta(t)=\frac{e^{-\eta t}}{(t+z)^\rho}\in\mathcal{S}.$$

Then, for $k=0,1,2,\ldots$ and $n=1,2,3,\ldots$, the following identities hold:

$$\lim_{\eta\to 0}\langle f,\varphi_\eta\rangle=\int_0^\infty\frac{f(t)}{(t+z)^\rho}dt,$$

$$\lim_{\eta\to 0}\langle\delta,\varphi_\eta^{(k)}\rangle=\frac{(-1)^k(\rho)_k}{z^{k+\rho}},$$

where $(\rho)_k$ denotes the Pochhammer's symbol,

$$\lim_{\eta\to 0}\langle\log(t),\varphi_\eta^{(k+1)}\rangle=\frac{(-1)^{k+1}}{z^{k+\rho}}(\rho)_k\big(\log(z)-\gamma-\psi(k+\rho)\big),$$

where γ is the Euler constant and ψ the digamma function and

$$\lim_{\eta\to 0}\langle f_{n,n},\varphi_\eta^{(n)}\rangle=(-1)^n(\rho)_n\int_0^\infty\frac{f_{n,n}(t)}{(t+z)^{n+\rho}}dt.$$

Proof. The first identity is trivial by using the dominated convergence theorem. The second one follows after a simple computation. On the other hand,

$$\langle \log(t), \varphi_\eta^{(k+1)} \rangle = (-1)^{k+1} \sum_{j=0}^{k+1} \binom{k+1}{j} \eta^j (\rho)_{k+1-j} \int_0^\infty \frac{e^{-\eta t}\log(t)}{(t+z)^{k+\rho+1-j}} dt.$$

For $j \leq k$ or $j = k+1$ and $\rho > 1$, the integrand of each integral on the right-hand side of the above equation is absolutely dominated by the integrable function $\log(t)(t+z)^{j-k-\rho-1}\ \forall\, \eta, t \geq 0$ and is therefore finite. For $j = k+1$ and $\rho \leq 1$, we divide the interval $[0,\infty)$ in the above integral at the point $t = 1$. On the interval $[0,1]$ the integral is finite for $\eta \geq 0$. In the interval $[1,\infty)$ we use the bound $\log(t) \leq \log(t+z)$, perform the change of variable $\eta t = u$, and divide again the resulting u-interval $[\eta,\infty)$ at the point $u = 1 - \eta z$ (assume $\eta \leq (1+z)^{-1}$). In the u-interval $[\eta, 1-\eta z]$ we use the bound $(u+\eta z)^\rho \geq u + \eta z$. After straightforward operations we obtain that the integral on the t-interval $[1,\infty)$ is $\mathcal{O}\left(\eta^{\rho-1}\log^2(\eta)\right)$ as $\eta \to 0$. Therefore,

$$\lim_{\eta\to 0} \langle \log(t), \varphi_\eta^{(k+1)} \rangle = (-1)^{k+1} (\rho)_{k+1} \int_0^\infty \frac{\log(t)}{(t+z)^{k+\rho+1}} dt.$$

Now using formula [16, p. 489, eq. (7)], we obtain the third identity. The fourth identity follows from the dominated convergence theorem, the local integrability of $f_{n,n}(t)$ on $[0,\infty)$, and the behavior $f_{n,n}(t) = \mathcal{O}(t^{-\alpha})$ as $t \to \infty$ [22, p. 296]. □

THEOREM 2.7. *Let $f(t)$ be a locally integrable function on $[0,\infty)$ and $\{A_k\}$ a sequence of complex numbers and let $f(t)$ have the following asymptotic expansion for large t and $n = 1, 2, 3, \ldots$:*

$$f(t) = \sum_{k=0}^{n-1} \frac{A_k}{t^{k+1}} + f_n(t), \tag{2.6}$$

where $f_n(t) = \mathcal{O}(t^{-n-1})$ as $t \to \infty$. Then, for $z, \rho > 0$ and $n = 1, 2, 3, \ldots$,

$$\int_0^\infty \frac{f(t)}{(t+z)^\rho} dt = \sum_{k=0}^{n-1} \frac{(-1)^k}{k! z^{k+\rho}} (\rho)_k \left[A_k\big(\log(z) - \gamma - \psi(k+\rho)\big) + B_k\right] + R_n(z), \tag{2.7}$$

where, for $k = 0, 1, 2, \ldots$, the coefficients B_k are given by

$$\begin{aligned} B_k &= A_k \sum_{j=1}^k \frac{1}{j} + \lim_{w\to k+1} \left\{ M[f;w] + \frac{A_k}{w-k-1} \right\} \\ &= A_k \sum_{j=1}^k \frac{1}{j} + \lim_{T\to\infty} \left\{ \int_0^T t^k f(t) dt - \sum_{j=0}^{k-1} A_j \frac{T^{k-j}}{k-j} - A_k \log(T) \right\}, \end{aligned} \tag{2.8}$$

empty sums being understood as zero. The remainder term is given by

$$R_n(z) = (\rho)_n \int_0^\infty \frac{f_{n,n}(t)}{(t+z)^{n+\rho}} dt, \tag{2.9}$$

where $f_{n,n}(t)$ is defined in (2.4).

Proof. From Lemmas 2.5 and 2.6 we obtain immediately (2.7), (2.9), and the first line in (2.8). Introducing

$$f_k(t)=f(t)-\sum_{j=0}^{k-1}\frac{A_j}{t^{j+1}}$$

in the second line of (2.5) and performing simple manipulations we obtain the second line in (2.8). □

A bound for the error term in the expansions given in Theorems 2.4 and 2.7 will be obtained in Propositions 2.10 and 2.11, respectively, when the function $f(t)$ has the form

$$f(t)=\prod_{k=1}^{m}\frac{1}{(t+x_k)^{\mu_k}}, \tag{2.10}$$

where $m\in\mathbb{N}$, $x_1,\dots,x_m$ are nonnegative parameters at least one different from zero, and $\mu_1,\dots,\mu_m>0$. Define

$$\mu=\sum_{k=1}^{m}\mu_k>0.$$

For $\mu\notin\mathbb{N}$, the asymptotic expansion of $f(t)$ in $t=\infty$ is given, for $n=1,2,3,\dots$, by

$$f(t)=\sum_{k=0}^{n-1}\frac{A_k}{t^{k+\mu-\lfloor\mu\rfloor}}+f_n(t), \tag{2.11}$$

where

$$A_0=A_1=\cdots=A_{\lfloor\mu\rfloor-1}=0 \qquad \text{if } \lfloor\mu\rfloor\ge 1,$$

$$A_{k+\lfloor\mu\rfloor}=\lim_{u\to 0}\frac{1}{k!}\frac{d^k}{du^k}\left(u^{-\mu}f(u^{-1})\right) \qquad \text{for } k=0,1,2,\dots, \tag{2.12}$$

and $f_n(t)=\mathcal{O}(t^{-n-\mu+\lfloor\mu\rfloor})$ as $t\to\infty$. Then, we have the following result.

LEMMA 2.8. *For $\mu\notin\mathbb{N}$ and $\forall\ t\in[0,\infty)$, the remainder term $f_n(t)$ and the coefficients A_n in the expansion* (2.11)–(2.12) *of the function $f(t)$ defined in* (2.10) *verify*

$$|f_n(t)|\le\frac{|A_n|}{t^{n+\mu-\lfloor\mu\rfloor}} \quad \text{for } n\ge\lfloor\mu\rfloor, \qquad |f_n(t)|\le\frac{|A_{n-1}|}{t^{n+\mu-\lfloor\mu\rfloor-1}} \quad \text{for } n\ge\lfloor\mu\rfloor+1, \tag{2.13}$$

and sign$(f_n(t))$ = sign(A_n) = sign$((-1)^{n-\lfloor\mu\rfloor})$ for $n\ge\lfloor\mu\rfloor$.

Proof. The Taylor expansion of $u^{-\mu}f(u^{-1})$ at $u=0$ is given by

$$u^{-\mu}f(u^{-1})\equiv\prod_{k=1}^{m}(1+x_ku)^{-\mu_k}=\sum_{k=0}^{n-\lfloor\mu\rfloor-1}A_{k+\lfloor\mu\rfloor}u^k+u^{-\mu}f_n(u^{-1}).$$

Applying the binomial formula for the derivative of a product we realize that the n-esim u-derivative of $u^{-\mu}f(u^{-1})$ has the same sign as $(-1)^n$ $\forall\ u\in[0,\infty)$. Then,

$\text{sign}(A_n) = \text{sign}(-1)^{n-\lfloor\mu\rfloor}$ for $n \geq \lfloor\mu\rfloor$ and, by the Lagrange formula for the remainder $u^{-\mu}f_n(u^{-1})$, we obtain that $\text{sign}(f_n(t)) = \text{sign}(-1)^{n-\lfloor\mu\rfloor}$ for $n \geq \lfloor\mu\rfloor$ and $\forall\, t \in [0,\infty)$. Therefore, two consecutive error terms $f_n(t)$ and $f_{n+1}(t)$ in the expansion of $f(t)$ have opposite sign. After applying the error test (see, for example, [15, p. 68] or [22, p. 38]) we obtain the first inequality in (2.13). The second inequality follows from the first one and

$$f_n(t) = f_{n-1}(t) - \frac{A_{n-1}}{t^{n+\mu-\lfloor\mu\rfloor-1}}. \qquad \square$$

On the other hand, for $\mu \in \mathbb{N}$, the asymptotic expansion in $t = \infty$ of the function $f(t)$ defined in (2.10) is given, for $n = 1, 2, 3, \ldots$, by

$$f(t) = \sum_{k=0}^{n-1} \frac{A_k}{t^{k+1}} + f_n(t), \tag{2.14}$$

where

$$A_0 = A_1 = \cdots = A_{\mu-2} = 0 \qquad \text{if } \mu \geq 2,$$

$$A_{k+\mu-1} = \lim_{u\to 0} \frac{1}{k!}\frac{d^k}{du^k}\left(u^{-\mu}f(u^{-1})\right) \qquad \text{for } k = 0, 1, 2, \ldots, \tag{2.15}$$

and $f_n(t) = \mathcal{O}(t^{-n-1})$ as $t \to \infty$. Then, we have the following lemma.

LEMMA 2.9. *For $\mu \in \mathbb{N}$ and $\forall\, t \in [0,\infty)$, the remainder term $f_n(t)$ and the coefficients A_n in the expansion* (2.14)–(2.15) *of the function $f(t)$ defined in* (2.10) *verify*

$$|f_n(t)| \leq \frac{|A_n|}{t^{n+1}} \qquad \text{for } n \geq \mu - 1, \qquad |f_n(t)| \leq \frac{|A_{n-1}|}{t^n} \qquad \text{for } n \geq \mu, \tag{2.16}$$

and sign$(f_n(t))$ = sign(A_n) = sign$((-1)^{n-\mu+1})$ for $n \geq \mu - 1$.

Proof. The proof is similar to the proof of Lemma 2.8 replacing $\lfloor\mu\rfloor$ by $\mu - 1$. $\square$

PROPOSITION 2.10. *If the function $f(t)$ of Theorem* 2.4 *has the form* (2.10) *with $\mu \notin \mathbb{N}$ then, $\forall\, p > 0$ and $n \geq \lfloor\mu\rfloor$, the error term $R_n(p)$ in the expansion* (2.2) *satisfies*

$$0 \leq (-1)^{\lfloor\mu\rfloor} R_n(p) \leq \frac{\pi|A_n|}{|\sin(\pi\mu)|p^{n+\mu-\lfloor\mu\rfloor}}, \tag{2.17}$$

providing the expansion (2.2) *of an asymptotic character for large p.*

Proof. The parameter α in Theorem 2.4 equals $\mu - \lfloor\mu\rfloor$ in Lemma 2.8. Using now $\text{sign}(f_n(u)) = \text{sign}((-1)^{n-\lfloor\mu\rfloor})\ \forall\ u \in [0,\infty)$ in (2.4) and (2.3) we obtain $(-1)^{\lfloor\mu\rfloor}R_n(p) \geq 0$. Introducing the first bound of (2.13) on the right-hand side of (2.4) and performing the change of variable $u = tv$ we obtain

$$|f_{n,n}(t)| \leq \frac{\Gamma(\mu - \lfloor\mu\rfloor)}{\Gamma(n+\mu-\lfloor\mu\rfloor)}\frac{|A_n|}{t^{\mu-\lfloor\mu\rfloor}} \qquad \forall \quad t \in [0,\infty).$$

Introducing this bound in (2.3) and after the change of variable $t = pu$ we obtain (2.17). $\square$

PROPOSITION 2.11. *If the function $f(t)$ of Theorem 2.7 has the form (2.10) with $\mu \in \mathbb{N}$ then, $\forall\ z > 0$ and $n \geq \mu$, the error term $R_n(z)$ in the expansion (2.7) satisfies the bounds*

$$0 \leq -(-1)^\mu R_n(z) \leq \frac{\pi\Gamma(n+\rho-1/2)}{\Gamma(\rho)\Gamma(n+1/2)} \frac{\bar{A}_n}{z^{n+\rho-1/2}}, \tag{2.18}$$

where $\bar{A}_n = \max\{|A_n|, |A_{n-1}|\}$ and

$$|R_n(z)| \leq \left[na|A_{n-1}| + |A_n|\big(S_n(z,a,\rho) + T_n(z,a,\rho)\big)\right] \frac{(\rho)_n}{n! z^{n+\rho}}, \tag{2.19}$$

where a is an arbitrary positive number,

$$S_n(z,a,\rho) = \min\left\{\frac{nz\left[(a+z)^{n+\rho-1} - z^{n+\rho-1}\right]}{a(n+\rho-1)(a+z)^{n+\rho-1}}, \psi(n+1)+\gamma\right\}, \tag{2.20}$$

and

$$\begin{aligned} T_n(z,a,\rho) &= \frac{z^{n+\rho}}{(n+\rho)(a+z)^{n+\rho}} F\left(n+\rho, 1; n+\rho+1; \frac{z}{a+z}\right) \\ &\leq \left(\frac{z}{a+z}\right)^\rho \left(\log\left(1+\frac{z}{a}\right) - \sum_{k=1}^{n-1} \frac{z^k}{k(z+a)^k}\right), \end{aligned} \tag{2.21}$$

where $F(b,c;d;z)$ is the hypergeometric function. For large z and fixed n, the optimum value for a is given by

$$a = \frac{|A_n|}{n|A_{n-1}|}. \tag{2.22}$$

Any of these bounds provides the expansion (2.7) of an asymptotic character for large z.

Proof. From Lemma 2.9, $\mathrm{sign}(f_n(u)) = \mathrm{sign}((-1)^{n-\mu+1})\ \forall\ u \in [0,\infty)$. Introducing this in (2.4) and (2.9) we obtain $(-1)^\mu R_n(z) \leq 0$. To obtain the bound (2.19) we divide the integral on the right-hand side of (2.4) by a fixed point $u = a \geq t$ and use the second bound of (2.16) in the integral over $[t,a]$ and the first bound of (2.13) in the integral over $[a,\infty)$. Using $u - t \leq u$ in the integral over $[t,a]$ we obtain

$$\begin{aligned} |f_{n,n}(t)| &\leq \frac{1}{(n-1)!}\left[|A_{n-1}|\log\left(\frac{a}{t}\right) + \frac{|A_n|}{nt}\left(1 - \left(1-\frac{t}{a}\right)^n\right)\right] \\ &\leq \frac{1}{(n-1)!}\left[|A_{n-1}|\log\left(\frac{a}{t}\right) + \frac{|A_n|}{a}\right] \qquad \forall\ t \in [0,a], \qquad a > 0. \end{aligned} \tag{2.23}$$

On the other hand, $\forall\ t \in [0,\infty)$ we introduce the first bound of (2.13) on the right-hand side of (2.4) and perform the change of variable $u = tv$. We obtain

$$|f_{n,n}(t)| \leq \frac{|A_n|}{n!}\frac{1}{t} \qquad \forall \quad t \in [0,\infty). \tag{2.24}$$

We divide the integral on the right-hand side of (2.9) at the point $t = a$ and use the bound (2.24) in the integral over $[a,\infty)$. Now, if we use the second bound of (2.23) in the integral over $[0,a]$ we obtain

$$|R_n(z)| \le \frac{(\rho)_n}{n!}\left[\frac{|A_n|S_n(z,a,\rho)}{z^{n+\rho}} + n|A_{n-1}|\int_0^a \frac{\log(a/t)}{(t+z)^{n+\rho}}dt \right. \\ \left. +|A_n|\int_a^\infty \frac{dt}{t(t+z)^{n+\rho}}\right], \tag{2.25}$$

where $S_n(z,a,\rho)$ is given by the first quantity between the brackets in (2.20). If instead of this, we use the first bound of (2.23) in the integral over $[0,a]$, expand $(1-t/a)^n$, and use the bound $(t+z)\ge z$ and [1, eq. (6.3.6)] we obtain again (2.25), but with $S_n(z,a,\rho)$ replaced by $\psi(n+1)+\gamma$.

A bound for the first integral on the right-hand side of (2.25) is given by $a/z^{n+\rho}$. After the change of variable $t=a/u$ in the second integral and using [18, eqs. (5.4)–(5.5)], we obtain (2.19) with $T_n(z,a,\rho)$ given by the right-hand side of the first line in (2.21). If, instead of computing exactly the second integral in (2.25), we use the bound $(t+z)^\rho \ge (a+z)^\rho$ $\forall$ $t\ge a$ and the equality [16, p. 31, eq. (4)], we obtain the second line in (2.21).

Finally, if we get rid of irrelevant terms for large z, the right-hand side of (2.19), as a function of a, has a minimum for a given in (2.22).

For obtaining the second inequality in (2.18), using Lemma 2.9 we have, for $n\ge\mu$, $|f_n(t)|\le|A_n|t^{-n-1/2}$ if $t\ge1$ and $|f_n(t)|\le|A_{n-1}|t^{-n-1/2}$ if $t\le1$. Therefore, $|f_n(t)|\le\bar A_n t^{-n-1/2}$ $\forall$ $t\in[0,\infty)$ and $n\ge\mu$. Then, $f_n(t)$ satisfies the first bound of (2.13) with μ replaced by $1/2$ and $|A_n|$ by $\bar A_n$. Repeating now the calculations of the proof of Proposition 2.10 we obtain the second inequality in (2.18). □

Remark 2.12. For large n and fixed z, the bound (2.19) (with a given in (2.22)) *contains an extra asymptotic factor* $\log(n)$ with respect to the bound (2.18), whereas for large z and fixed n, it *contains an extra asymptotic factor* $\log(z)/\sqrt{z}$. Therefore, (2.19) is more suitable for large z and (2.18) is more suitable for large n.

3. Expansions of the symmetric standard EIs.

Convergent expansions of R_F, R_D, and R_J for large values of one of their parameters are obtained as corollaries of Theorem 2.4 or 2.7. Error bounds for the remainder terms in these expansions follow from Propositions 2.10 and 2.11. We derive the explicit expansions and error bounds for the remainders in the following subsections.

3.1. Expansion of $R_F(x,y,z)$ for large z.

Corollary 3.1. *A uniformly convergent expansion of* $R_F(x,y,z)$ *for* $0\le x<y\le z$ *is given, for* $n=1,2,3,\dots,$ *by*

$$R_F(x,y,z) = \frac{1}{2\sqrt{z}}\sum_{k=0}^{n-1}\frac{(-1)^k}{k!z^k}\left(\frac{1}{2}\right)_k\left[A_k^F(x,y)\left(\log(z)-\gamma-\psi\left(k+\frac{1}{2}\right)\right)\right. \\ \left. +B_k^F(x,y)\right] + R_n^F(x,y,z), \tag{3.1}$$

where, for $k=0,1,2,\dots,$

$$A_k^F(x,y) = (-1)^k\sum_{j=0}^{k}\frac{(1/2)_j(1/2)_{k-j}}{j!(k-j)!}x^j y^{k-j} \tag{3.2}$$

and coefficients $B_k^F(x,y)$ *are given by the recursion*

$$B_{k+2}^F = \frac{A_{k+2}^F}{k+1} + \frac{xyA_k^F + (x+y)A_{k+1}^F + 2A_{k+2}^F}{k+2} + \frac{2k+3}{2k+4}(x+y)\left[\frac{A_{k+1}^F}{k+1} - B_{k+1}^F\right] - \frac{k+1}{k+2}xyB_k^F, \tag{3.3}$$

empty sums being understood as zero and

$$B_0^F = -2\log\left(\frac{\sqrt{x}+\sqrt{y}}{2}\right), \qquad B_1^F = (x+y)\log\left(\frac{\sqrt{x}+\sqrt{y}}{2}\right) - \sqrt{xy}. \tag{3.4}$$

For $n = 1, 2, 3, \ldots,$ *the remainder* $R_n^F(x,y,z)$ *is positive and a bound for* $R_n^F(x,y,z)$ *is given by the right-hand side of* (2.18) *or* (2.19) *putting* $\rho \equiv 1/2$ *and* $A_n \equiv A_n^F(x,y)$ *given above. In particular, two error bounds are given by*

$$\begin{aligned} R_n^F(x,y,z) &\leq \frac{1}{2n!}\left(\frac{1}{2}\right)_n \left[1 + \psi(n+1) + \gamma + \log\left(1 + \frac{nz|A_{n-1}^F|}{|A_n^F|}\right)\right]\frac{|A_n^F|}{z^n\sqrt{z}}, \\ R_n^F(x,y,z) &\leq \frac{\sqrt{\pi}(n-1)!}{\Gamma(n+1/2)}\frac{\bar{A}_n^F}{z^n}, \end{aligned} \tag{3.5}$$

where $\bar{A}_n^F = \max\{|A_n^F|, |A_{n-1}^F|\}$.

Proof. The integral $2R_F(x,y,z)$ has the form considered in Theorem 2.7 with

$$f(t) \equiv f^F(t) = \frac{1}{\sqrt{(t+x)(t+y)}} = \sum_{k=0}^{n-1}\frac{A_k^F}{t^{k+1}} + f_n^F(t), \tag{3.6}$$

where $f_n^F(t) = \mathcal{O}(t^{-n-1})$ as $t \to \infty$ and $\rho = 1/2$. Therefore, the asymptotic expansion of $2R_F(x,y,z)$ for large z follows from (2.7) in Theorem 2.7. Coefficients $A_k \equiv A_k^F(x,y)$ in (2.6) are trivially given by (3.2).

For calculating $B_k \equiv B_k^F(x,y)$ we consider the second line in (2.8). Define, for $k = 0, 1, 2, \ldots,$

$$I_k^F(x,y,T) \equiv \int_0^T t^k f^F(t)dt \equiv \int_0^T \frac{t^k}{\sqrt{(t+x)(t+y)}}dt$$

and

$$\sigma_k^F(x,y) \equiv \lim_{T\to\infty}\left\{I_k^F(x,y,T) - \sum_{j=0}^{k-1}A_j^F\frac{T^{k-j}}{k-j} - A_k^F\log(T)\right\}. \tag{3.7}$$

Integrals $I_k^F(x,y,T)$ satisfy the recursion

$$I_{k+2}^F = \frac{1}{2(k+2)}\left[2T^{k+1}\sqrt{(T+x)(T+y)} - (2k+3)(x+y)I_{k+1}^F - 2(k+1)xyI_k^F\right]. \tag{3.8}$$

On the other hand, from the differential equation $2(t+x)(t+y)(f^F)' + (2t+x+y)f^F = 0$, we obtain, for $k = 0, 1, 2, \ldots,$

$$2(k+2)A_{k+2}^F + (2k+3)(x+y)A_{k+1}^F + 2(k+1)xyA_k^F = 0. \tag{3.9}$$

Now we substitute $I^F_{k+2}(x,y,T)$ in the definition (3.7) of $\sigma^F_{k+2}(x,y)$ with the right-hand side of (3.8), expand the term $\sqrt{(T+x)(T+y)}$ in inverse powers of T, and use recursion (3.9). We obtain

$$2(k+2)\sigma^F_{k+2}=2xyA^F_k+2(x+y)A^F_{k+1}+2A^F_{k+2}-(2k+3)(x+y)\sigma^F_{k+1}-2(k+1)xy\sigma^F_k,$$

from which (3.3) follows easily by using the second lines in (2.8) and (3.9). Integrals $I^F_0(x,y,T)$ and $I^F_1(x,y,T)$ may be calculated by using formula [16, p. 53, eqs. (3) and (8)]. Then, from the second line in (2.8) and using $A^F_0=1$ and $A^F_1=-(x+y)/2$ we obtain (3.4).

Function $f^F(t)$ satisfies the conditions of Proposition 2.11 with $\mu=1$. Therefore, $R^F_n(x,y,z)\geq 0$ and the bounds (2.18) and (2.19) hold for $2R^F_n(x,y,z)$ setting $\rho\equiv 1/2$ and $A_n\equiv A^F_n(x,y)$ given in (3.2). In particular, introducing (2.22) in (2.19) we obtain the first line of (3.5).

Introducing the bound $|A^F_n|\leq y^n$ in the second line of (3.5) we obtain, for $n\geq 1$,

$$R^F_n(x,y,z)\leq C(y,z)\frac{y^n}{z^n\sqrt{n}}, \tag{3.10}$$

where $C(y,z)$ is independent of n. Therefore, expansion (3.1) is uniformly convergent for $y\leq z$. □

Remark 3.2. An alternative (and explicit) expression for the coefficients $B^F_k(x,y)$ can be obtained from the first line in (2.8). Using the equality [16, p. 303, eq. (24)] and the reflection formula of the gamma function [1, eq. (6.1.17)] we have, for $w\notin\mathbb{Z}$,

$$M[f^F;w]=\frac{\pi}{\sin(\pi w)}\frac{x^w}{\sqrt{xy}}F\left(w,\frac{1}{2};1;1-\frac{x}{y}\right)\qquad \text{if } \ y>x>0,$$

$$M[f^F;w]=\frac{\sqrt{\pi}}{\sin(\pi w)}\frac{\Gamma(w-1/2)}{\Gamma(w)}y^{w-1}\qquad \text{if } \ y>x=0 \ \text{ and } \ \frac{3}{2}-w\notin\mathbb{N}.$$

Subtracting the pole $-A_k/(w-k-1)$ and taking the limit $w\to k+1$ we obtain

$$B^F_k(x,y)=A^F_k(x,y)\sum_{j=1}^{k}\frac{1}{j}-(-1)^kC^F_k(x,y), \tag{3.11}$$

where

$$C^F_k(0,y)=\frac{(1/2)_ky^k}{k!}\left(\psi\left(k+\frac{1}{2}\right)-\psi(k+1)+\log(y)\right) \tag{3.12}$$

and

(3.13)
$$C^F_k(x,y)=x^k\sqrt{\frac{x}{y}}\left(\log(x)F\left(k+1,\frac{1}{2};1;1-\frac{x}{y}\right)+F'\left(k+1,\frac{1}{2};1;1-\frac{x}{y}\right)\right)$$

for $x>0$, where $F'(a,b;c;z)$ denotes the first derivative of $F(a,b;c;z)$ with respect to the argument a.

Remark 3.3. Formulas (3.11)–(3.13) provide coefficients B^F_k in expansion (3.1) of an explicit expression which may be useful for analytical purposes. On the other hand, recurrence (3.3)–(3.4) involves only elementary functions and may be more appropriate for numerical computations. Similar comments can be made about the coefficients B^D_k in the expansion of $R_D(x,z,y)$ for large z given in section 3.3.

Table 3.1 shows a numerical example of the approximation supplied by expansion (3.1).

Table 3.1

Second, third, and sixth columns represent $R_F(1,2,z)$, approximation (3.1) for $n=1$, and approximation (3.1) for $n=2$, respectively. Fourth and seventh columns represent the respective relative error $-R_n^F(1,2,z)/R_F(1,2,z)$. Fifth and last columns represent the respective error bounds given by (2.19).

z	$R_F(1,2,z)$	1st-order approx.	Relative error	Relative error bound	2nd-order approx.	Relative error	Relative error bound
10	0.5537947453	0.5237406385	−0.0543	0.0725789476	0.5504844438	−0.00598	0.0100213601
20	0.4609268635	0.4478367679	−0.0284	0.0366674515	0.4601982389	−0.00158	0.0024544658
50	0.3522219102	0.3480283802	−0.0119	0.0148009679	0.3521274856	−0.000268	0.0003865126
100	0.2824793637	0.2807505868	−0.00612	0.0074304273	0.2824597696	−0.0000694	0.0000958172
200	0.2237272736	0.2230270972	−0.00313	0.0037243942	0.2237232837	−0.0000178	0.0000237907

3.2. Expansion of $R_D(x,y,z)$ for large z.

Corollary 3.4. *A uniformly convergent expansion of $R_D(x,y,z)$ for $0 \le x < y < z$ is given, for $n = 1,2,3,\ldots,$ by*

$$\begin{aligned} R_D(x,y,z) = \frac{3}{2\sqrt{z^3}} \sum_{k=0}^{n-1} \frac{(-1)^k}{k! z^k} \left(\frac{3}{2}\right)_k \Bigg[A_k^F(x,y) \left(\log(z) - \gamma - \psi\left(k+\frac{3}{2}\right)\right) \\ + B_k^F(x,y) \Bigg] + R_n^D(x,y,z), \end{aligned} \tag{3.14}$$

where $A_k^F(x,y)$ and $B_k^F(x,y)$ are given in (3.2) *and* (3.3)–(3.4) *(or* (3.11)–(3.13)*), respectively.*

For $n = 1,2,3,\ldots,$ the remainder term $R_n^D(x,y,z)$ is positive and a bound for $(2/3)R_n^D(x,y,z)$ is given by the right-hand side of (2.18) *or* (2.19) *putting $\rho \equiv 3/2$ and $A_n \equiv A_n^F(x,y)$ given in* (3.2)*. In particular, two error bounds are given by*

$$\begin{aligned} R_n^D(x,y,z) &\le \frac{3}{2n!}\left(\frac{3}{2}\right)_n \left[1+\psi(n+1)+\gamma+\log\left(1+\frac{nz|A_{n-1}^F|}{|A_n^F|}\right)\right]\frac{|A_n^F|}{z^n\sqrt{z^3}}, \\ R_n^D(x,y,z) &\le \frac{2\sqrt{\pi}n!}{\Gamma(n+1/2)}\frac{\bar{A}_n^F}{z^{n+1}}. \end{aligned}$$

Proof. The integral $(2/3)R_D(x,y,z)$ has the form considered in Theorem 2.7 with $f(t) \equiv f^F(t)$ given in (3.6) and $\rho = 3/2$. The remaining proof follows as in Corollary 3.1 except that, in this case, (3.10) reads

$$R_n^D(x,y,z) \le C(y,z)\frac{y^n\sqrt{n}}{z^n}$$

and convergence of expansion (3.14) is restricted to $y < z$. □

Table 3.2 shows a numerical example of the approximation supplied by expansion (3.14).

3.3. Expansion of $R_D(x,z,y)$ for large z.

Corollary 3.5. *A uniformly convergent expansion of $R_D(x,z,y)$ for $0 \le x < y < z$ or $0 \le y < x < z$ is given, for $n = 1,2,3,\ldots,$ by*

$$\begin{aligned} R_D(x,z,y) = \frac{3}{2\sqrt{z}} \sum_{k=0}^{n-1} \frac{(-1)^k}{k! z^k} \left(\frac{1}{2}\right)_k \Bigg[A_k^D(x,y) \left(\log(z) - \gamma - \psi\left(k+\frac{1}{2}\right)\right) \\ + B_k^D(x,y) \Bigg] + \bar{R}_n^D(x,z,y), \end{aligned} \tag{3.15}$$

TABLE 3.2

Second, third, and sixth columns represent $R_D(1,2,z)$, approximation (3.14) for $n = 1$, and approximation (3.14) for $n = 2$, respectively. Fourth and seventh columns represent the respective relative error $-R_n^D(1,2,z)/R_D(1,2,z)$. Fifth and last columns represent the respective error bounds given by (2.19).

z	$R_D(1,2,z)$	1st-order approx.	Relative error	Relative error bound	2nd-order approx.	Relative error	Relative error bound
10	0.0835011776	0.0622538617	−0.254	0.3565374834	0.0792081617	−0.0514	0.0913314070
20	0.0384213534	0.0336344955	−0.125	0.1669454033	0.0379393692	−0.0125	0.0203200273
50	0.0130325135	0.0123964214	−0.0488	0.0624511404	0.0130069811	−0.00196	0.0029107969
100	0.0055565283	0.0054225176	−0.0241	0.0299926153	0.0055538440	−0.000483	0.0006835181
200	0.0023123734	0.0022847462	−0.0120	0.0145049853	0.0023120972	−0.000119	0.0001625233

where $A_0^D(x,y) = 0$ and, for $k = 1,2,3,\dots$,

$$A_k^D(x,y) = (-1)^{k-1}\sum_{j=0}^{k-1}\frac{(1/2)_j(3/2)_{k-j-1}}{j!(k-j-1)!}x^j y^{k-j-1}. \tag{3.16}$$

Coefficients $B_k^D(x,y)$ for $k = 0,1,2,\dots$ are given by the recurrence

$$\begin{aligned} B_{k+2}^D = {} & \frac{A_{k+2}^D}{k+2} + \frac{xyA_k^D + (x+y)A_{k+1}^D + 2A_{k+2}^D}{k+1} \\ & + \left[\frac{x-y}{2k+2} - x - y\right]\left[B_{k+1}^D - \frac{A_{k+1}^D}{k+1}\right] - xyB_k^D, \end{aligned} \tag{3.17}$$

empty sums being understood as zero and

$$B_0^D = \frac{2}{y-x}\left(1-\sqrt{\frac{x}{y}}\right), \qquad B_1^D = 1 + \frac{2y}{x-y}\left(1-\sqrt{\frac{x}{y}}\right) - 2\log\left(\frac{\sqrt{x}+\sqrt{y}}{2}\right). \tag{3.18}$$

For $n = 1,2,3,\dots$, the remainder term $\bar{R}_n^D(x,z,y)$ is negative,

$$|\bar{R}_1^D(x,z,y)| \le \frac{3}{2\sqrt{z}(r-z)} + \frac{3}{4\sqrt{(z-r)^3}}\log\left[\frac{\sqrt{z}+\sqrt{z-r}}{\sqrt{z}-\sqrt{z-r}}\right] \tag{3.19}$$

if $r \equiv \min\{x,y\} > 0$,

$$|\bar{R}_1^D(x,z,y)| \le \frac{3}{y\sqrt{z}} - \frac{3\pi}{4\sqrt{z^3}}F\left(\frac{3}{2},\frac{3}{2};2;1-\frac{y}{z}\right) \qquad \text{if } \; y > x \ge 0, \tag{3.20}$$

and, for $n = 2,3,4,\dots$, a bound for $|\bar{R}_n^D(x,z,y)|$ is given by the right-hand side of (2.18) or (2.19) setting $\rho \equiv 1/2$ and $A_n \equiv A_n^D(x,y)$ given above. In particular, two error bounds are given, for $n \ge 2$, by

$$\begin{aligned} |\bar{R}_n^D(x,z,y)| &\le \frac{3}{2n!}\left(\frac{1}{2}\right)_n\left[1+\psi(n)+\gamma+\log\left(1+\frac{nz|A_{n-1}^D|}{|A_n^D|}\right)\right]\frac{|A_n^D|}{z^n\sqrt{z}}, \\ |\bar{R}_n^D(x,z,y)| &\le \frac{\sqrt{\pi}(n-1)!}{\Gamma(n+1/2)}\frac{\bar{A}_n^D}{z^n}, \end{aligned} \tag{3.21}$$

where $\bar{A}_n^D = \max\{|A_n^D|, |A_{n-1}^D|\}$.

Proof. The integral $(2/3)R_{D}(x,z,y)$ has the form required in Theorem 2.7 with

$$f(t)\equiv f^{D}(t)=\frac{1}{\sqrt{(t+x)(t+y)^3}}=\sum_{k=0}^{n-1}\frac{A_k^{D}}{t^{k+1}}+f_n^{D}(t),$$

where $f_n^{D}(t)=\mathcal{O}(t^{-n-1})$ as $t\to\infty$ and $\rho=1/2$. Therefore, the asymptotic expansion of $(2/3)R_{D}(x,z,y)$ follows from (2.7) in Theorem 2.7. Trivially, the coefficients $A_k\equiv A_k^{D}(x,y)$ in (2.6) satisfy $A_0^{D}=0$ and, for $k=1,2,3,\dots$, they are given by (3.16).

Recurrence (3.17)–(3.18) for $B_k^{D}(x,y)$ follows from the second line in (2.8). Its derivation follows the pattern of derivation of (3.3)–(3.4) in Corollary 3.1. We define, for $k=0,1,2,\dots$,

$$I_k^{D}(x,y,T)\equiv\int_0^T f^{D}(t)dt\equiv\int_0^T\frac{t^k}{\sqrt{(t+x)(t+y)^3}}dt$$

and

$$\sigma_k^{D}(x,y)\equiv\lim_{T\to\infty}\left\{I_k^{D}(x,y,T)-\sum_{j=0}^{k-1}A_j^{D}\frac{T^{k-j}}{k-j}-A_k^{D}\log(T)\right\}. \tag{3.22}$$

Integrals $I_k^{D}(x,y,T)$ satisfy the recurrence

$$I_{k+2}^{D}=\frac{T^{k+1}}{k+1}\sqrt{\frac{T+x}{T+y}}+\left(\frac{x-y}{2k+2}-x-y\right)I_{k+1}^{D}-xyI_k^{D}. \tag{3.23}$$

On the other hand, from the differential equation $2(t+x)(t+y)(f^{D})'+(4t+3x+y)f^{D}=0$, we obtain, for $k=0,1,2,\dots$,

$$2(k+1)A_{k+2}^{D}+(2(k+1)(x+y)+y-x)A_{k+1}^{D}+2(k+1)xyA_k^{D}=0. \tag{3.24}$$

Now we substitute $I_{k+2}^{D}(x,y,T)$ in the definition (3.22) of $\sigma_{k+2}^{D}(x,y)$ by the right-hand side of (3.23), expand the term $\sqrt{(T+x)/(T+y)}$ in inverse powers of T, and use the recurrence (3.24). We obtain

$$2(k+1)\sigma_{k+2}^{D}=2xyA_k^{D}+2(x+y)A_{k+1}^{D}+2A_{k+2}^{D}+(x-y-2(k+1)(x+y))\sigma_{k+1}^{D}-2(k+1)xy\sigma_k^{D},$$

from which (3.17) follows easily by using the second lines in (2.8) and (3.24). Integrals $I_0^{D}(x,y,T)$ and $I_1^{D}(x,y,T)$ can be calculated by using formulas [16, p. 53, eqs. (6) and (8)]. Then, from the second line in (2.8), $A_0^{D}=0$, and $A_1^{D}=1$ we obtain (3.18).

Function $f^{D}(t)$ has the form required in Proposition 2.11 with $\mu=2$. Therefore, $\bar{R}_n^{D}(x,z,y)\le 0$ and the bounds (2.18) and (2.19) hold for $(2/3)\bar{R}_n^{D}(x,z,y)$ setting $\rho=1/2$ and $A_n\equiv A_n^{D}(x,y)$ given in (3.16) for $n=2,3,4,\dots$. In particular, the first line of (3.21) follows after introducing (2.22) in inequality (2.19). On the other hand, $A_0^{D}=0$ means $f_1^{D}(t)=f^{D}(t)$. Introducing the bounds $f^{D}(u)\le(u+r)^{-2}$ if $r=\min\{x,y\}>0$ or $f^{D}(u)\le\sqrt{u}(u+y)^{-3/2}$ if $y>x\ge 0$ in the definition (2.4) of $f_{1,1}(t)$ and using [16, p. 52, eq. (6)] or [16, p. 53, eq. (4)], respectively, we obtain (3.19) and (3.20).

Using the second line of (3.21) and $|A_{n+1}^{D}(x,y)|\le(n+1)s^n$, where $s=\max\{x,y\}$, we obtain, for $n\ge 1$,

$$|\bar{R}_n^{D}(x,z,y)|\le C(s,z)\frac{s^n\sqrt{n}}{z^n},$$

Table 3.3

Second, third, and sixth columns represent $R_D(1,z,2)$, approximation (3.15) for $n=1$, and approximation (3.15) for $n=2$, respectively. Fourth and seventh columns represent the respective relative error $-\bar{R}_n^D(1,z,2)/R_D(1,z,2)$. Fifth and last columns represent the respective error bounds given by (2.19).

z	$R_D(1,z,2)$	1st-order approx.	Relative error	Relative error bound	2nd-order approx.	Relative error	Relative error bound
10	0.2390532443	0.2778629050	0.162	0.2021314366	0.2508051822	0.0492	0.0870637346
20	0.1784332265	0.1964787444	0.101	0.1221700356	0.1811001790	0.0149	0.0238026222
50	0.1180725686	0.1242640688	0.0524	0.0612672806	0.1184298191	0.00303	0.0043925602
100	0.0852097498	0.0878679657	0.0312	0.0357104912	0.0852853865	0.000888	0.0012264633
200	0.0610194998	0.0621320343	0.0182	0.0205219425	0.0610351563	0.000257	0.0003410757

where $C(s,z)$ is independent of n. Therefore, expansion (3.15) is uniformly convergent for $s<z$. $\square$

Remark 3.6. An alternative (and explicit) expression for the coefficients $B_k^D(x,y)$ can be obtained by using the first equality in (2.8) and

$$M[f^D;w]=\frac{\pi(1-w)x^w}{\sin(\pi w)\sqrt{xy^3}}F\left(w,\frac{3}{2};2;1-\frac{x}{y}\right)\qquad \text{if } \ x,y>0,\quad w\notin\mathbb{Z},$$

or

$$M[f^D;w]=\frac{2\sqrt{\pi}y^{w-2}}{\sin(\pi(w-1))}\frac{\Gamma(w-1/2)}{\Gamma(w-1)}\qquad \text{if } \ y>x=0,\quad w\notin\mathbb{Z},\quad \frac{3}{2}-w\notin\mathbb{N}.$$

The derivation of these formulas is similar to the derivation of $M[f^F;w]$ in Remark 2.12. Subtracting the pole $-A_k^D/(w-k-1)$ and taking the limit $w\to k+1$ we obtain, for $k=0,1,2,\dots$,

$$B_k^D(x,y)=A_k^D(x,y)\sum_{j=1}^{k}\frac{1}{j}+(-1)^kC_k^D(x,y), \tag{3.25}$$

where $C_0^F(0,y)=2/y$,

$$C_k^F(0,y)=\frac{2k(1/2)_k y^{k-1}}{k!}\big(\psi(k+1/2)-\psi(k)+\log(y)\big)\qquad \text{for } \ k=1,2,3,\dots,$$

and, for $k=0,1,2,\dots$ and $x,y>0$,

$$C_k^F(x,y)=x^k\sqrt{\frac{x}{y^3}}\left[(1+k\log(x))F\left(k+1,\frac{3}{2};2;1-\frac{x}{y}\right)+kF'\left(k+1,\frac{3}{2};2;1-\frac{x}{y}\right)\right].$$

Table 3.3 shows a numerical example of the approximation supplied by expansion (3.15).

3.4. Expansion of $R_J(x,y,z,p)$ for large z.

Corollary 3.7. *A uniformly convergent expansion of $R_J(x,y,z,p)$ for $0<p<z$ and $0\le x<y<z$ is given, for $n=1,2,3,\dots$, by*

$$\begin{aligned}R_J(x,y,z,p)=\frac{3}{2\sqrt{z}}\sum_{k=0}^{n-1}\frac{(-1)^k}{k!z^k}\left(\frac{1}{2}\right)_k&\left[A_k^J(x,y,p)\left(\log(z)-\gamma-\psi\left(k+\frac{1}{2}\right)\right)\right.\\ &\left.+B_k^J(x,y,p)\right]+R_n^J(x,y,z,p),\end{aligned} \tag{3.26}$$

where $A_0^J(x,y,p)=0$ *and, for* $k=1,2,3,\dots,$

$$A_k^J(x,y,p)=\sum_{j=0}^{k-1}(-p)^{k-j-1}A_j^F(x,y), \tag{3.27}$$

where $A_j^F(x,y)$ *are given in* (3.2). *Coefficients* $B_k^J(x,y,p)$ *are given by the recurrence*

$$B_{k+1}^J=B_k^F-pB_k^J+A_{k+1}^J\sum_{j=1}^{k+1}\frac{1}{j}+(pA_k^J-A_k^F)\sum_{j=1}^{k}\frac{1}{j}, \qquad k=0,1,2,\dots, \tag{3.28}$$

where $B_k^F(x,y)$ *are given by* (3.3)–(3.4) *(or* (3.11)–(3.13)*), empty sums must be understood as zero, and*

$$B_0^J=\frac{2}{\sqrt{|(p-x)(p-y)|}}\log\left[\frac{\sqrt{p(p-x)}+\sqrt{p(p-y)}}{\sqrt{y(p-x)}+\sqrt{x(p-y)}}\right] \qquad \text{if}\ \ p>x,y, \tag{3.29}$$

$$B_0^J=\frac{2}{\sqrt{|(p-x)(p-y)|}}\log\left[\frac{\sqrt{x(y-p)}+\sqrt{y(x-p)}}{\sqrt{p(x-p)}+\sqrt{p(y-p)}}\right] \qquad \text{if}\ \ p<x,y, \tag{3.30}$$

$$\begin{aligned} B_0^J=\frac{1}{\sqrt{|(p-x)(p-y)|}}&\left[\sin^{-1}\left(\frac{x+y-2p}{y-x}\right)\right.\\ &\left.-\sin^{-1}\left(\frac{2xy-p(x+y)}{p(y-x)}\right)\right] \qquad \text{if}\ \ x<p<y. \end{aligned} \tag{3.31}$$

For $n=1,2,3,\dots,$ *the remainder term* $R_n^J(x,y,z,p)$ *is negative,*

$$|R_1^J(x,y,z)|\le\frac{3}{2\sqrt{z}(r-z)}+\frac{3}{4\sqrt{(z-r)^3}}\log\left[\frac{\sqrt{z}+\sqrt{z-r}}{\sqrt{z}-\sqrt{z-r}}\right] \tag{3.32}$$

if $r\equiv\min\{x,y,p\}>0,$

$$|R_1^J(x,y,z)|\le\frac{3}{r\sqrt{z}}-\frac{3\pi}{4\sqrt{z^3}}F\left(\frac{3}{2},\frac{3}{2};2;1-\frac{r}{z}\right) \tag{3.33}$$

if $r\equiv\min\{y,p\}>0$ *or* $r\equiv\min\{x,p\}>0,$ *and, for* $n=2,3,4,\dots,$ *a bound for* $(2/3)|R_n^J(x,y,z,p)|$ *is given by the right-hand side of* (2.18) *or* (2.19) *putting* $\rho\equiv1/2$ *and* $A_n\equiv A_n^J(x,y,p)$ *given above. In particular, two error bounds are given, for* $n\ge2$, *by*

$$\begin{aligned} |R_n^J(x,y,z,p)|&\le\frac{3}{2n!}\left(\frac{1}{2}\right)_n\left[1+\psi(n+1)+\gamma+\log\left(1+\frac{nz|A_{n-1}^J|}{|A_n^J|}\right)\right]\frac{|A_n^J|}{z^n\sqrt{z}},\\ |R_n^J(x,y,z,p)|&\le\frac{\sqrt{\pi}(n-1)!}{\Gamma(n+1/2)}\frac{\bar{A}_n^J}{z^n}, \end{aligned} \tag{3.34}$$

where $\bar{A}_n^J=\max\{|A_n^J|,|A_{n-1}^J|\}$.

Proof. The integral $(2/3)R_J(x,y,z,p)$ has the form required in Theorem 2.7 with

$$f(t)\equiv f^J(t)=\frac{1}{\sqrt{(t+x)(t+y)(t+p)}}=\sum_{k=0}^{n-1}\frac{A_k^J}{t^{k+1}}+f_n^J(t),$$

Table 3.4

Second, third, and sixth columns represent $R_J(1,2,z,3)$, approximation (3.26) *for $n=1$, and approximation* (3.26) *for $n=2$, respectively. Fourth and seventh columns represent the respective relative error $-R_n^J(1,2,z,3)/R_J(1,2,z,3)$. Fifth and last columns represent the respective error bounds given in* (2.19).

z	$R_J(1,2,z,3)$	1st-order approx.	Relative error	Relative error bound	2nd-order approx.	Relative error	Relative error bound
10	0.1877070842	0.2227576125	0.187	0.2574232928	0.2013272410	0.0726	0.1357463320
20	0.1409922070	0.1575134184	0.117	0.1546127555	0.1441244221	0.0222	0.0367400044
50	0.0938633074	0.0996202328	0.0613	0.0770693616	0.0942893087	0.00454	0.0067564415
100	0.0679464537	0.0704421421	0.0367	0.0447835297	0.0680375155	0.00134	0.0018883978
200	0.0487571525	0.0498101164	0.0216	0.0256831788	0.0487761541	0.000390	0.0005263141

where $f_n^J(t)=\mathcal{O}(t^{-n-1})$ as $t\to\infty$ and $\rho=1/2$. The asymptotic expansion of $(2/3)R_J(x,y,z,p)$ for large z follows from (2.7) in Theorem 2.7. Trivially, $A_0^J=0$ and, for $k=1,2,3,\ldots$ coefficients A_k^J are given by (3.27).

Recurrence (3.28) for $B_k^J(x,y,p)$ follows from the second line in (2.8). We define, for $k=0,1,2,\ldots,$

$$I_k^J(x,y,p,T)\equiv\int_0^T t^k f^D(t)dt\equiv\int_0^T\frac{t^k}{\sqrt{(t+x)(t+y)}(t+p)}dt$$

and

$$\sigma_k^J(x,y,p)\equiv\lim_{T\to\infty}\left\{I_k^J(x,y,p,T)-\sum_{j=0}^{k-1}A_j^J\frac{T^{k-j}}{k-j}-A_k^J\log(T)\right\}. \tag{3.35}$$

Integrals $I_k^J(x,y,p,T)$ satisfy the recurrence $I_{k+1}^J=I_k^F-pI_k^J$. Then, (3.28) follows easily by using (2.8), (3.7), (3.27), and (3.35). Integral $I_0^J(x,y,p,T)$ may be calculated by using [16, pp. 53, 54, eqs. (8)–(11)]. Then, from the second line in (2.8) and $A_0^J=0$ we obtain (3.29)–(3.31).

Function $f^J(t)$ has the form required in Proposition 2.11 with $\mu=2$. Therefore, $R_n^J(x,y,z,p)\le 0$ and the bounds (2.18) and (2.19) hold for $(2/3)R_n^J(x,y,z,p)$ setting $\rho=1/2$ and $A_n\equiv A_n^J(x,y,p)$ given in (3.27) for $n=2,3,4,\ldots$. In particular, the first line of (3.34) follows after introducing (2.22) in inequality (2.19). On the other hand, $A_0^J=0$ means $f_1^J(t)=f^J(t)$. Then, after a similar calculation to the one used for deriving (3.19) and (3.20), we obtain (3.32) and (3.33).

Using the second line of (3.34) and the bound $|A_{n+1}^J(x,y,p)|\le(n+1)s^n$, where $s=\max\{x,y,p\}$, we obtain, for $n\ge 1$,

$$|R_n^J(x,y,z,p)|\le C(s,z)\frac{s^n\sqrt{n}}{z^n},$$

where $C(s,z)$ is independent of n. Therefore, expansion (3.26) is uniformly convergent for $s<z$. □

Table 3.4 shows a numerical example of the approximation supplied by expansion (3.26).

3.5. Expansion of $R_J(x,y,z,p)$ for large p.

Corollary 3.8. *A uniformly convergent expansion of $R_J(x,y,z,p)$ for $0\le x<y<z<p$ is given, for $n=1,2,3,\ldots,$ by*

$$R_J(x,y,z,p)=\frac{3}{2}\sum_{k=0}^{n-1}\frac{D_k^J(x,y,z)}{p^{k+1}}-\frac{3\pi}{2\sqrt{p}}\sum_{k=0}^{n-1}\frac{C_k^J(x,y,z)}{p^k}+\bar{R}_n^J(x,y,z,p), \tag{3.36}$$

where $C_0^J(x,y,z)=0$ *and, for* $k=1,2,3,\ldots,$

$$C_k^J(x,y,z)=\sum_{j=0}^{k-1}\sum_{s=0}^{k-j-1}\frac{(1/2)_j(1/2)_s(1/2)_{k-j-s-1}}{j!s!(k-j-s-1)!}x^jy^sz^{k-j-s-1}. \tag{3.37}$$

Coefficients $D_k^J(x,y,z)$ *are given by the recurrence*

$$\begin{aligned}D_{k+3}^J=&\frac{1}{2k+5}\left[2(k+2)(x+y+z)D_{k+2}^J-(2k+3)(xy+xz+yz)D_{k+1}^J\right.\\&\left.+2(k+1)xyzD_k^J\right],\end{aligned} \tag{3.38}$$

$$D_0^J(x,y,z)=2R_F(x,y,z), \tag{3.39}$$

$$D_1^J(x,y,z)=\frac{2}{3}(z-x)(y-z)R_D(x,y,z)+2zR_F(x,y,z)+2\sqrt{\frac{xy}{z}}, \tag{3.40}$$

and

$$D_2^J(x,y,z)=\frac{2}{3}(x+y+z)D_1^J(x,y,z)-\frac{2}{3}(xy+xz+yz)R_F(x,y,z)-\frac{2}{3}\sqrt{xyz}. \tag{3.41}$$

For $n=1,2,3,\ldots,$ *the remainder term* $\bar{R}_n^J(x,y,z,p)$ *is negative and a bound is given by*

$$|\bar{R}_n^J(x,y,z,p)|\le\frac{3\pi C_n^J(x,y,z)}{2p^n\sqrt{p}}. \tag{3.42}$$

Proof. The integral $(2/3)R^J(x,y,z,p)$ has the form required in Theorem 2.4 with

$$f(t)\equiv\bar{f}^J(t)=\frac{1}{\sqrt{(t+x)(t+y)(t+z)}}=-\sum_{k=0}^{n-1}\frac{(-1)^kC_k^J}{t^{k+1/2}}+\bar{f}_n(t), \tag{3.43}$$

where $\bar{f}_n(t)=\mathcal{O}(t^{-n-1/2})$ as $t\to\infty$ and $\alpha=1/2$. Therefore, the asymptotic expansion of $(2/3)R^J(x,y,z,p)$ for large p is given by (2.2). Coefficients $A_k\equiv-(-1)^kC_k^J(x,y,z)$ are trivially given for $k=1,2,3,\ldots,$ by (3.37) and $C_0^J(x,y,z)=0$. Coefficients $D_k^J(x,y,z)\equiv(-1)^kM[\bar{f}^J;k+1]$ in (2.2) represent the analytic continuation of the Mellin transform of $\bar{f}^J(t)$ evaluated in $k+1$. The Mellin transform of the function $\bar{f}^J(t)$ given in (3.43) is defined by (2.1) for $0<\mathrm{Re}(w)<3/2$. We divide the integration path in this formula at $t=1$ and, in the integral over $[1,\infty)$, we substitute $\bar{f}^J(t)$ by the right-hand side of (3.43) with n replaced by $n+1$:

$$M[\bar{f}^J;w]=\int_0^1t^{w-1}\bar{f}^J(t)dt+\sum_{k=0}^{n}\frac{(-1)^kC_k^J(x,y,z)}{w-k-1/2}+\int_1^\infty t^{w-1}\bar{f}_{n+1}^J(t)dt.$$

The first integral is an analytic function of w for $\mathrm{Re}(w)>0$. The second integral is analytic for $\mathrm{Re}(w)<n+3/2$. Therefore, this formula gives the analytic continuation of $M[\bar{f}^J;w]$ to the strip $0<\mathrm{Re}(w)<n+3/2$ (which has simple poles at $w=k+1/2$,

$k=0,1,2,\ldots,n$). We evaluate $M[\bar f^J;w]$ above at the point $w=k+1$ and replace $\bar f^J_{n+1}(t)$ in the last integral by $\bar f^J(t)+\sum_{k=0}^n(-1)^kC_k^Jt^{-k-1/2}$. After straightforward operations we obtain

$$D_k^J(x,y,z)=(-1)^k\lim_{T\to\infty}\left\{\int_0^T t^k\bar f^J(t)dt+\sum_{j=0}^k\frac{(-1)^jC_j^JT^{k-j+1/2}}{k-j+1/2}\right\}. \tag{3.44}$$

We define, for $s>0$,

$$\alpha_k(x,y,z,s,T)\equiv\sqrt{s}\int_0^T\frac{t^k}{\sqrt{(t+x)(t+y)(t+z)(t+s)}}dt$$

and

$$I_k(x,y,z,T)\equiv\int_0^T\frac{t^k}{\sqrt{(t+x)(t+y)(t+z)}}dt=\lim_{s\to\infty}\alpha_k(x,y,z,s,T).$$

Integrals $\alpha_k(x,y,z,s,T)$ satisfy the recurrence

$$\begin{aligned}2T^{k+1}\sqrt{s(T+x)(T+y)(T+z)(T+s)}&=2(k+3)\alpha_{k+4}\\&+(2k+5)(x+y+z+s)\alpha_{k+3}+2(k+2)(xy+xz+xs+yz+ys+zs)\alpha_{k+2}\\&+(2k+3)(xyz+xys+xzs+yzs)\alpha_{k+1}+2(k+1)xyzs\alpha_k.\end{aligned}$$

Taking the limit $s\to\infty$ we obtain that the integrals $I_k(x,y,z,T)$ satisfy the recurrence

$$\begin{aligned}I_{k+3}=\frac{1}{2k+5}\Big[&2T^{k+1}\sqrt{(T+x)(T+y)(T+z)}-2(k+2)(x+y+z)I_{k+2}\\&-(2k+3)(xy+xz+yz)I_{k+1}-2(k+1)xyzI_k\Big].\end{aligned}\tag{3.45}$$

On the other hand, from the differential equation $2(t+x)(t+y)(t+z)(\bar f^J)'+(3t^2+2(x+y+z)t+xy+xz+yz)\bar f^J=0$, we obtain, for $k=0,1,2,\ldots$,

$$2(k+1)C_{k+2}^J-(2k+1)(x+y+z)C_{k+1}^J+2k(xy+xz+yz)C_k^J-(2k-1)xyzC_{k-1}^J=0. \tag{3.46}$$

If we expand the term $\sqrt{(T+x)(T+y)(T+z)}$ in (3.45) in inverse powers of T and use the recurrence (3.46) and the definition (3.44), we obtain the recurrence (3.38). Using (3.44) we see that D_0^J is trivially given by (3.39). Integrating $I_1(x,y,z)$ by parts in (3.44) and using [16, p. 71, eq. (10)] and [18, eq. (12.33)] we obtain (3.40). Equation (3.41) follows after straightforward operations.

Function $f^J(t)$ satisfies the conditions of Proposition 2.10 with $\mu=3/2$. Therefore, $\bar R_n^J(x,y,z,p)\le 0$ and the bound (2.17) holds for $(2/3)\bar R_n^J(x,y,z,p)$ setting $A_n\equiv(-1)^nC_n^J$, and (3.42) follows. Using (3.42) and the bound $|C_{n+1}^J(x,y,z)|\le(3/2)_nz^n/n!$, we obtain, for $n\ge1$,

$$|\bar R_n^J(x,y,z,p)|\le C(z,p)\frac{\sqrt{n}z^n}{p^n},$$

where $C(z,p)$ is independent of n. Therefore, expansion (3.36) is uniformly convergent for $z<p$. □

Table 3.5

Second, third, and sixth columns represent $R_J(1,2,3,p)$, approximation (3.36) for $n = 2$, and approximation (3.36) for $n = 3$, respectively. Fourth and seventh columns represent the respective relative error $-\bar{R}_n^J(1,2,3,p)/R_J(1,2,3,p)$. Fifth and last columns represent the respective error bounds given by (3.42).

				Relative			Relative
		2nd-order	Relative	error	3rd-order	Relative	error
z	$R_J(1,2,3,p)$	approx.	error	bound	approx.	error	bound
10	0.1237859612	0.1531757825	0.237	0.3611528060	0.1316685706	0.0637	0.0963074149
20	0.0716068743	0.0773834863	0.0807	0.1103653337	0.0723803740	0.0108	0.0147153778
50	0.0330037076	0.0336525403	0.0197	0.0242311844	0.0330384089	0.00105	0.0012923298
100	0.0178156797	0.0179370973	0.00682	0.0079352385	0.0178189241	0.000182	0.0002116063
200	0.0094259946	0.0094483849	0.00238	0.0026513081	0.0094262936	0.0000317	0.0000353507

Table 3.5 shows a numerical example of the approximation supplied by expansion (3.36).

Remark 3.9. A bound for the n-esim remainder term in any of the expansions given in Corollaries 3.1–3.8 has the form $C(s,z)\sqrt{n}(s/z)^n$ for $n \geq 1$, where z is the asymptotic variable, s is a bound for the remaining variables, and $C(s,z)$ is independent of n. Therefore, the convergence rate of these expansions increases for decreasing value of the quotient s/z.

4. Conclusions. Following Wong's proposal [21, Example 1], the distributional approach has been used in Theorem 2.7 for deriving an alternative proof for the asymptotic expansion of the generalized Stieltjes transforms (see [21, Theorem 2 and Example 1]). Using this result we have derived convergent expansions of $R_F(x,y,z)$, $R_D(x,y,z)$, $R_D(x,z,y)$, and $R_J(x,y,z,p)$ for $x,y,p < z$ in Corollaries 3.1–3.7, respectively. On the other hand, using the asymptotic expansion of the Stieltjes transforms [22, chap. 6, sect. 2, Theorem 1], we have obtained a convergent expansion of $R_J(x,y,z,p)$ for $x,y,z < p$ in Corollary 3.8. Functions $f(t)$ in the integrand of R_F, R_D, and R_J (and, in general, functions $f(t)$ given in (2.10)) are to a special kind of function: the remainder terms in their asymptotic expansions in inverse powers of t satisfy the error test. This fundamental property is used in Propositions 2.10 and 2.11 for deriving an error bound for the remainder in the asymptotic expansions given in Theorems 2.4 and 2.7 at any order of the approximation. In particular, it has been derived for the expansions of R_F, R_D, and R_J in Corollaries 3.1–3.8. These bounds have been obtained from the error test and, as numerical computations show (see Tables 3.1–3.5), they exhibit a remarkable accuracy. Moreover, these bounds show that the expansions are convergent when the asymptotic variable is greater than the remaining ones and that the convergence rate increases as this difference between the asymptotic variable and the remaining ones increases.

Expansions given in Corollaries 3.1–3.8 are generalizations of the corresponding first-order approximations given by Carlson and Gustafson [10]. Nevertheless, the complete expansion of the first EI given in Corollary 3.1 was already obtained by Carlson and Gustafson [4] and, as well as in Corollary 3.1, the coefficients of the expansion are given by a recurrence [4, eq. (1.7)]. Complete expansions for R_D and R_J for the asymptotic parameters considered in Corollaries 3.4–3.8 were also obtained by Carlson and Gustafson [3]. But a recurrence for the calculation of the coefficients is not given in [3] and error bounds supplied there are not quite satisfactory. The advantage of the approach presented here is that it supplies a simple algorithm for the calculation of the coefficients of these expansions and more accurate error bounds at any order of the approximation. This algorithm is explicitly given in Corollaries 3.1–3.8. Moreover, coefficients of the expansions of $R_F(x,y,z)$, $R_D(x,y,z)$, $R_D(x,z,y)$, and $R_J(x,y,z,p)$

for large z are given in terms of elementary functions, whereas coefficients of the expansion of $R_J(x,y,z,p)$ for large p are given in terms of $R_F(x,y,z)$ and $R_D(x,y,z)$.

For large z, the error bounds supplied in Corollaries 3.1–3.7 are slightly more accurate than the error bounds given in [10] for the first order approximation of $R_D(x,z,y)$, $R_F(x,y,z)$, and $R_J(x,y,z,p)$ and slightly less accurate for the first-order approximation of $R_D(x,y,z)$ and the second-order approximation of $R_F(x,y,z)$. When considering first-order approximations to $R_J(x,y,z,p)$ for large p, a comparison between the error bounds given in Corollary 3.8 and the error bounds given in [10] is more complicated because they are concerned with different approximations. On the other hand, at any order of the approximation, error bounds given in [4, eqs. (1.20) and (3.40)] are more accurate for large values of n than the error bound given in Corollary 3.1 and less accurate for small n.

The distributional approach should succeed for deriving complete uniform asymptotic expansions of symmetric EI too. This challenge is postponed for further investigations.

Acknowledgments. This work was stimulated by conversations with Nico Temme, whose continued interest has been quite encouraging. I am grateful to the anonymous referees for their careful examination of the paper and their helpful suggestions.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists*, Springer-Verlag, New York, 1971.

[3] B. C. Carlson, *The hypergeometric function and the R-function near their branch points*, Rend. Sem. Mat. Univ. Politec. Torino, special issue, (1985), pp. 63–89.

[4] B. C. Carlson and J. L. Gustafson, *Asymptotic expansion of the first elliptic integral*, SIAM J. Math. Anal., 16 (1985), pp. 1072–1092.

[5] B. C. Carlson, *A table of elliptic integrals of the second kind*, Math. Comp., 49 (1987), pp. 595–606.

[6] B. C. Carlson, *A table of elliptic integrals of the third kind*, Math. Comp., 51 (1988), pp. 267–280.

[7] B. C. Carlson, *A table of elliptic integrals: Cubic cases*, Math. Comp., 53 (1989), pp. 327–333.

[8] B. C. Carlson, *A table of elliptic integrals: One quadratic factor*, Math. Comp., 56 (1991), pp. 267–280.

[9] B. C. Carlson, *A table of elliptic integrals: Two quadratic factors*, Math. Comp., 59 (1992), pp. 165–180.

[10] B. C. Carlson and J. L. Gustafson, *Asymptotic approximations for symmetric elliptic integrals*, SIAM J. Math. Anal., 25 (1994), pp. 288–303.

[11] M. Ghandehari and D. Logothetti, *How elliptic integrals K and E arise from circles and points in the Minkowski plane*, J. Geom., 50 (1994), pp. 63–72.

[12] J. L. Gustafson, *Asymptotic Formulas for Elliptic Integrals*, Ph.D. thesis, Iowa State University, Ames, IA, 1982.

[13] E. L. Kaplan, *Auxiliary table for the incomplete elliptic integrals*, J. Math. Phys., 27 (1948), pp. 11–36.

[14] A. M. Legendre, *Traité des fonctions elliptiques, Vol.* I, Imprimerie de Huzard-Courcier, Paris, 1925.

[15] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[16] A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, *Integrals and Series, Vol.* 1, Gordon and Breach, New York, 1990.

[17] M. Razpet, *An application of elliptic integrals*, J. Math. Anal. Appl., 168 (1992), pp. 425–429.

[18] N. M. Temme, *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*, Wiley, New York, 1996.

[19] A. M. Urbina et al., *Elliptic integrals and limit cycles*, Bull. Austral. Math. Soc., 48 (1993), pp. 195–200.

[20] W. S. Weiglhofer, *Electromagnetic depolarization dyadics and elliptic integrals*, J. Phys. A., 31 (1998), pp. 7191–7196.

[21] R. Wong, *Explicit error terms for asymptotic expansions of Mellin convolutions*, J. Math. Anal. Appl., 72 (1979), pp. 740–756.

[22] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 776–790

A SYSTEM OF DEGENERATE PARABOLIC EQUATIONS FROM PLASMA PHYSICS: THE LARGE TIME BEHAVIOR*

M. BERTSCH† AND S. KAMIN‡

Abstract. The authors study the large time behavior of solutions of a strongly coupled system of degenerate parabolic equations describing the heat and mass transfer in a one-dimensional plasma. The model was introduced by Hyman and Rosenau, and the existence of solutions was studied in our earlier paper. The results of this paper describe how the large time behavior depends on some of the parameters in the equations.

Key words. nonlinear system, plasma physics, large time behavior, degenerate parabolic equations

AMS subject classifications. 35K65, 35K45, 35B40

PII. S0036141098336613

1. Introduction.

We consider the following system of nonlinear partial differential equations:

$$\text{(I)}\qquad \begin{cases} \rho_t = (\varphi_1(T)\rho^{a_1}\rho_x)_x & \text{in } Q := \mathbb{R}\times\mathbb{R}^+, \\ (\rho T)_t = (\varphi_2(T)\rho^{a_2+1}T_x)_x + (T\varphi_1(T)\rho^{a_1}\rho_x)_x & \text{in } Q, \\ \rho(x,0) = \rho_0(x) & \text{for } x\in\mathbb{R}, \\ T(x,0) = T_0(x) & \text{for } x\in\mathcal{P}_0\ , \end{cases}$$

where

$$\mathcal{P}_0 = \{x\in\mathbb{R} : \rho_0(x) > 0\}\ . \tag{1.1}$$

Here x is a one-dimensional spatial coordinate, t indicates time, $\rho(x,t)$ and $T(x,t)$ are nonnegative functions to be determined, ρ_0 and T_0 are given nonnegative and continuous functions, a_1 and a_2 are positive constants, and φ_1 and φ_2 are smooth positive functions.

System (I) was introduced by Rosenau and Hyman [12], [13] to study the effect of nonlinearly coupled mass and heat diffusion in a plasma which slowly diffuses in a strong magnetic field. In this context ρ and T denote the density and ionic temperature of the plasma, and φ_1 and φ_2 are power-type functions of T. The last term in the differential equation for the temperature T represents the heat transport due to mass diffusion.

In an earlier paper [5] we have constructed a solution of Problem (I). In the present paper we describe the large time behavior of this solution in the case where

*Received by the editors April 1, 1998; accepted for publication (in revised form) September 23, 1999; published electronically March 21, 2000. Partial support was obtained through the project Nonlinear Parabolic Equations, funded by TMR ERB FMRX CT98 0201 of the European Community.

http://www.siam.org/journals/sima/31-4/33661.html

†Dipartimento di Matematica, Università di Roma Tor Vergata, Via della Ricerca Scientifica, 00133 Roma, Italy (bertsch@vaxiac.iac.rm.cnr.it). This author was supported by the Consiglio Nazionale delle Ricerche under a grant in the programme Mathematical Methods in Fluid Dynamics and Molecular Dynamics.

‡School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel (kamin@math.tau.ac.il). This author was supported by the Israel Science Foundation under grant 505/95-1.

ρ_0 is integrable in $\mathbb{R}$, i.e., the case of finite total mass. As we shall see below, this behavior is quite different in the cases $a_1 \le a_2$ and $a_1 > a_2$; in other words, the large time behavior depends on the parameter values of a_1 and a_2.

To explain our results we introduce the precise hypotheses on the data:

(H1) $\varphi_i \in C^2(\mathbb{R}^+), \varphi_i > 0$ in $\mathbb{R}^+$, $a_i \in \mathbb{R}^+$ $(i = 1, 2)$;

(H2) $\rho_0 \in C(\mathbb{R}) \cap L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, $\rho_0 \ge 0$ and $\rho_0 \not\equiv 0$ in $\mathbb{R}$, $\mathcal{P}_0 \subseteq \mathbb{R}$ is defined by (1.1);

(H3) $T_0 \in C(\mathcal{P}_0)$ and $0 < \nu_0 \le T_0 \le \nu_1$ in $\mathcal{P}_0$ for some constants ν_0 and ν_1.

Since $a_1 > 0$, the equation for ρ degenerates at points where ρ vanishes (if $\varphi_1 \equiv 1$, we obtain the well-known porous medium equation; see, for example, [1]). The equation does not always possess classical solutions. Therefore we have to introduce solutions in a weak sense, and in addition define T only in the set

$$\mathcal{P} = \{(x,t) \in \overline{Q} : \rho(x,t) > 0\}. \tag{1.2}$$

Definition 1.1. *A pair* (ρ, T) *is called a solution of Problem* (I) *if*

(i) $\rho \in C(\overline{Q}) \cap L^\infty(Q)$, $T \in C(\mathcal{P}) \cap L^\infty(\mathcal{P})$, *where* $\mathcal{P}$ *is defined by* (1.2)*;*

(ii) $\rho \ge 0$ *in* Q, $T \ge \mu$ *in* $\mathcal{P}$ *for some* $\mu > 0$*;*

(iii) $(\rho^{a_1+1})_x \in L^2_{\mathrm{loc}}(\overline{Q})$, $T_x \in L^2_{\mathrm{loc}}(\mathcal{P})$, $\rho^{a_2+1}T_x \in L^2(S)$ *for any bounded measurable set* $S \subseteq \mathcal{P}$*;*

(iv) *for any* $\psi \in C^{1,1}(\overline{Q})$ *with compact support*

$$\int_{\mathcal{P}_0} \rho_0(x)\psi(x,0)dx + \int\int_{\mathcal{P}} \Big(\rho\psi_t - \frac{1}{a_1+1}\varphi_1(T)(\rho^{a_1+1})_x\psi_x\Big)dxdt = 0$$

and

$$\int_{\mathcal{P}_0} \rho_0(x)T_0(x)\psi(x,0)dx + \int\int_{\mathcal{P}} \Big(\rho T\psi_t - \rho^{a_2+1}\varphi_2(T)T_x\psi_x \\ - \frac{1}{a_1+1}T\varphi_1(T)(\rho^{a_1+1})_x\psi_x\Big)dxdt = 0,$$

where the set $\mathcal{P}_0$ *is defined by* (1.1)*.*

In [5] we have constructed a solution as the limit of classical solutions $(\rho_\epsilon, T_\epsilon)$ of the following approximate problem:

$$(\mathrm{I}_\epsilon) \quad \begin{cases} \rho_t = (\varphi_1(T)\rho^{a_1}\rho_x)_x & \text{in } Q_\epsilon = (-L_\epsilon, L_\epsilon) \times \mathbb{R}^+, \\ (\rho T)_t = (\varphi_2(T)\rho^{a_2+1}T_x)_x + (T\varphi_1(T)\rho^{a_1}\rho_x)_x & \text{in } Q_\epsilon, \\ \rho(x,0) = \rho_{0\epsilon}(x), T(x,0) = T_{0\epsilon}(x) & \text{for } x \in (-L_\epsilon, L_\epsilon), \\ \rho_x(\pm L_\epsilon, t) = T_x(\pm L_\epsilon, t) = 0 & \text{for } t > 0. \end{cases}$$

Here $\epsilon \in (0, \rho^*)$, where

$$\rho^* = \sup\{\rho_0(x), x \in \mathbb{R}\},$$

$L_\epsilon > 0$, $\rho_{0\epsilon}$, $T_{0\epsilon} \in C^\infty([-L_\epsilon, L_\epsilon])$, $\rho'_{0\epsilon}(\pm L_\epsilon) = T'_{0\epsilon}(\pm L_\epsilon) = 0$,

$$0 < \epsilon \le \rho_{0\epsilon} \le \rho^*, \quad \nu_0 \le T_{0\epsilon} \le \nu_1 \text{ in } (-L_\epsilon, L_\epsilon),$$

and $\rho_{0\epsilon} \to \rho_0$ in $C_{\mathrm{loc}}(\mathbb{R})$, $T_{0\epsilon} \to T_0$ in $C_{\mathrm{loc}}(\mathcal{P}_0)$ and $L_\epsilon \to \infty$ as $\epsilon \to 0^+$. Since $\rho_0 \in L^1(\mathbb{R})$ we assume that L_ϵ and $\rho_{0\epsilon}$ are chosen such that, for sufficiently small values of $\epsilon > 0$,

$$\int_{-L_\epsilon}^{L_\epsilon} \rho_{0\epsilon}(x)dx = M := \int_{-\infty}^{\infty} \rho_0(x)dx. \tag{1.3}$$

In [5] we have shown that there exist $\rho \in C(\overline{Q})$, $T \in C(\mathcal{P})$ and a sequence $\epsilon_n \to 0^+$ such that

$$\rho_n := \rho_{\epsilon_n} \to \rho \text{ in } C_{\text{loc}}(\overline{Q}) \text{ and } T_n := T_{\epsilon_n} \to T \text{ in } C_{\text{loc}}(\mathcal{P}) \text{ as } n \to \infty, \tag{1.4}$$

and we have proved that (ρ, T) is a solution of Problem (I). In addition ρ and T are classical solutions of the system in the set $\mathcal{P}$, and

$$\rho_n \to \rho \text{ and } T_n \to T \text{ in } C^{2,1}_{\text{loc}}(\mathcal{P}\backslash\{t=0\}) \text{ as } n \to \infty. \tag{1.5}$$

The uniqueness of the solution is not known.

Now we are ready to describe the large time behavior of ρ and T. First we observe that

$$\rho(x,t) \le \mathcal{C}\|\rho_0\|_{L^1(\mathbb{R})} t^{-1/(a_1+2)} \quad \text{for} \quad (x,t) \in Q,$$

where the constant $\mathcal{C}$ depends only on a_1 and min/max $\varphi(s)$ for $s \in [\nu_0, \nu_1]$. This decay rate is essentially a result for the equation for ρ, in which $\varphi_1(T(x,t))$ is considered as a given function (see section 2 for some results concerning the porous medium equation with nonconstant coefficients).

Our main result is the behavior of $\rho(x,t)$ and $T(x,t)$ as $t \to \infty$.

Theorem 1.2 (large time behavior). *Let hypotheses* (H1)–(H3) *be satisfied and let* (ρ, T) *be the solution of Problem* (I) *defined by* (1.4). *Then there exist positive constants* η_0 *and* η_1 *and functions* $\overline{R} \in C(\mathbb{R}) \cap C^2((-\eta_0, \eta_1))$ *and* $\overline{T} \in C^1((-\eta_0, \eta_1))$ *such that*

(i) $\overline{R} > 0$ *in* $(-\eta_0, \eta_1)$, $\overline{R} = 0$ *in* $\mathbb{R}\backslash(-\eta_0, \eta_1)$, *and* $\nu_0 \le \overline{T} \le \nu_1$ *in* $(-\eta_0, \eta_1)$;

(ii) *for any* $C > 0$

$$t^{1/(a_1+2)}\rho(x,t) \to \overline{R}(xt^{-1/(a_1+2)}) \textit{ as } t \to \infty \tag{1.6}$$

uniformly with respect to x *satisfying* $-Ct^{1/(a_1+2)} \le x \le Ct^{1/(a_1+2)}$, *and for any* $\epsilon > 0$

$$T(x,t) \to \overline{T}(xt^{-1/(a_1+2)}) \textit{ as } t \to \infty \tag{1.7}$$

uniformly with respect to x *satisfying*

$$-(\eta_0 - \epsilon)t^{1/(a_1+2)} \le x \le (\eta_1 - \epsilon)t^{1/(a_1+2)};$$

(iii) *the function* $\overline{\rho}$ *defined by*

$$\overline{\rho}(x,t) = t^{-1/(a_1+2)}\overline{R}(xt^{-1/(a_1+2)})$$

is a (self-similar) solution of the equation

$$\rho_t = \Big(\varphi_1\big(\overline{T}(xt^{-1/(a_1+2)})\big)\rho^{a_1}\rho_x\Big)_x \tag{1.8}$$

with initial data Dirac's δ*-function multiplied by* $M = \|\rho_0\|_{L^1(\mathbb{R})}$:

$$\int_{\mathbb{R}} \psi(x)\overline{\rho}(x,t)dx \to M\psi(0) \textit{ as } t \to 0^+ \tag{1.9}$$

for all $\psi \in C(\mathbb{R})$ *with compact support;*

(iv) *if* $a_1 \geq a_2$, *then*

$$\overline{T}(\eta) = \text{const.} = \frac{\int_{\mathcal{P}_0} \rho_0(x) T_0(x) dx}{M} \ , \tag{1.10}$$

and if $a_1 < a_2$, *then there exist initial data* (ρ_0, T_0) *such that* $(\overline{T})_x \not\equiv 0$.

Given the function $\overline{T}(\eta)$, substitution of the self-similar solution $\overline{\rho}(x,t)$ in (1.8) yields the following ordinary differential equation for $\overline{R}(\eta)$ in the set where $\overline{R} > 0$:

$$\Big(\varphi_1(\overline{T}(\eta)) \overline{R}^{a_1} \overline{R}'\Big)' = -\frac{1}{a_1+2} (\eta \overline{R})'. \tag{1.11}$$

Integrating twice we obtain the formula

$$\overline{R}^{a_1}(\eta) = \left(C - \frac{a_1}{a_1+2} \int_0^\eta \frac{s}{\varphi_1(\overline{T}(s))} ds \right)_+ \ ,$$

where we have used the notation $a_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$. The integration constant C is uniquely determined by the initial mass M (condition (1.9)):

$$\int_{\mathbb{R}} \overline{R}(\eta) d\eta = M = \|\rho_0\|_{L^1(\mathbb{R})}.$$

In the case of the "classical" porous medium equation ($\varphi_1 \equiv 1$) $\overline{\rho}(x,t)$ is the Barenblatt solution [3]

$$\overline{\rho}(x,t) = t^{-1/(a_1+2)} \left(C - \frac{a_1 \eta^2}{2(a_1+2)} \right)_+^{1/a_1} \ , \quad \eta = x t^{-\frac{1}{a_1+2}}.$$

It is well known [10], [14] that the Barenblatt solution describes the large time behavior of solutions of the porous medium equation with finite total mass M.

The most interesting part of Theorem 1.2 seems to be part (iv). If $a_1 \geq a_2$, T becomes constant as $t \to \infty$, i.e., the temperature converges to its average value given by (1.10), and, for large values of t, ρ behaves as the Barenblatt solution with the same total mass. However, if $a_1 < a_2$ the temperature does not always reach its average value, and ρ behaves asymptotically as a solution of (1.11) which may be different from the Barenblatt solution. Physically we could give the following interpretation of this parameter dependence: if $a_1 < a_2$ the diffusivity $\varphi_2(T)\rho^{a_2}$ of the heat diffusion is, for large values of t, i.e., for small values of ρ, much smaller than the diffusivity $\varphi_1(T)\rho^{a_1}$ of the mass diffusion, and apparently the heat diffusion is too weak to make the temperature constant as $t \to \infty$. In the case of an initial boundary value problem Rosenau and Hyman [12] conjectured such phenomena studying separable solutions and presenting numerical evidence.

In the proof of Theorem 1.2 we decouple the two equations of the system as much as possible. For this purpose, we introduce the mass (or Lagrangian) variable

$$y = \int_{-\infty}^x \rho(s,t) ds \in [0, M] \quad \text{for} \quad t \geq 0, \tag{1.12}$$

where M is the total initial mass, defined by (1.3).

In [5] we have shown that the transformation $(x,t) \to (y,t)$ leads to the following set of equations for $\hat{\rho}(y,t) \equiv \rho(x,t)$ and $\hat{T}(y,t) \equiv T(x,t)$:

$$(1.13) \qquad \begin{cases} \hat{\rho}_t = \hat{\rho}^2(\varphi_1(\hat{T})\hat{\rho}^{a_1}\hat{\rho}_y)_y & \text{in } (0,M)\times\mathbb{R}^+, \\ \hat{T}_t = (\varphi_2(\hat{T})\hat{\rho}^{a_2+2}\hat{T}_y)_y & \text{in } (0,M)\times\mathbb{R}^+. \end{cases}$$

System (1.13) was used in our paper [5] to prove the existence of a solution of Problem (I). The advantage of the mass variable is that the second term in the equation for T, which represents the heat transport due to the diffusion of mass, disappears and this fact allows us to study the large time behavior of T as a function of y and t. Subsequently the information about the behavior of the coefficient $\varphi_1(T(x,t))$ in the porous medium equation for ρ is sufficient to determine the large time behavior of ρ.

The paper is organized as follows. In section 2 we list some properties of the porous medium equation with nonconstant coefficients. The essential part of the results we refer to in section 2 are obtained in our paper [7]. In fact our study of the porous medium equation with nonconstant coefficients was inspired by our work on the system. It is worth mentioning here that the main difficulty in this study is that we do not have much information about the smoothness of the coefficient $A(x,t)$ (see section 2) but only its upper and lower bounds. In section 3 we introduce the scaling of ρ, while in section 4 we prove the convergence of T as $t\to\infty$ in terms of the Lagrangian coordinates. In section 5, we translate the results of the former sections in terms of the original x-variable and prove Theorem 1.2, and in section 6, we obtain some estimates used in previous sections.

Finally we mention that in [8] the existence of a solution of the Cauchy and boundary value problem has been recently proved in the case of higher spatial dimension.

2. The porous medium equation with nonconstant coefficients. In this section we collect some results about the problem

$$(\text{II}) \qquad \begin{cases} u_t = (A(x,t)(u^m)_x)_x & \text{in } Q, \\ u(x,0) = u_0(x) \geq 0 & \text{for } x \in \mathbb{R}, \end{cases}$$

where the assumptions on the data are the following.

Assumption A1. $m>1$, $A \in L^\infty(Q)$, $0<\lambda\leq A(x,t)\leq\Lambda$ (a.e.) in Q for some constants λ and Λ, $u_0 \in L^\infty(\mathbb{R})\cap C(\mathbb{R})$ and $u_0 \geq 0$ in $\mathbb{R}$.

We observe that the equation for $\rho(x,t)$ in System (I) reduces to the one in Problem (II) if we set $A(x,t) = \frac{1}{m}\varphi(T(x,t))$, $m = a_1+1$.

As we already pointed out in the introduction we have the right to use only the information on the coefficient $A(x,t)$ which is supplied to us by the solution of the system. Thus in general this coefficient is not smooth, but on the other hand it is known by (1.4) and (1.5) that $\rho(x,t) = \lim \rho_{\epsilon_n}(x,t)$. Therefore we give the following definition.

Definition 2.1. *Let $A_n(x,t)$ be a sequence of functions ($n=1,2,\dots$) such that each A_n is defined in the domain $\Omega_n = (-L_n,L_n)\times\mathbb{R}^+$, $L_n\to\infty$ as $n\to\infty$. Moreover assume that A_n satisfies*

$$A_n(x,t) \in L^\infty(\Omega_n)\cap C^{2,2}(\overline{\Omega}_n), \quad 0\leq\lambda\leq A_n(x,t)\leq\Lambda \quad \text{in} \quad \Omega_n,$$

where λ and Λ do not depend on n. Suppose that $A_n\to A$ in $L^1_{\text{loc}}(Q)$. Let u_n be the *classical solution of the problem*

$$(\mathrm{II}_n) \qquad \begin{cases} \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} A_n \frac{\partial u^m}{\partial x} & \text{in } \Omega_n, \\ \frac{\partial u}{\partial x}(\pm L_n, t) = 0 & \text{for } t > 0, \\ u(x,0) = u_{0n}(x) & \text{for } x \in (-L_n, L_n), \end{cases}$$

where u_{0n} is a smooth positive function, $u_{0n} \to u_0$ in $C_{\mathrm{loc}}(\mathbb{R})$ as $n \to \infty$, and, if $u_0 \in L^1(\mathbb{R})$, then

$$\int_{-L_n}^{L_n} u_{on}(x)dx = \int_{-\infty}^{\infty} u_0(x)dx.$$

Suppose that

$$u_n(x,t) \to u(x,t) \quad \text{in} \quad C_{\mathrm{loc}}(Q) \quad \text{and} \quad \frac{\partial u_n^m}{\partial x} \to \frac{\partial u^m}{\partial x} \textit{ weakly in } L^2_{\mathrm{loc}}(Q).$$

Then the limit function u is said to be a solution of the Cauchy problem (II). *Obviously $u(x,t)$ satisfies the equation*

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} A \frac{\partial u^m}{\partial x}$$

in the sense of distributions and $u(x,0) = u_0(x)$.

The definition given above is equivalent to Definition 6.1 in [7]. Therefore we use the results of this paper below.

The following properties hold for a solution of Problem (II) in the sense of the above definition [7, section 6].

Proposition 2.2. *Let Assumption* A1 *be satisfied. Then for any compact set $K \subset \mathbb{R}$ the modulus of continuity of u in $K \times [0,T]$ depends only on $\sup_Q u, T, \lambda, \Lambda$ and the modulus of continuity of u_0 in an open neighborhood $\widetilde{K}$ of K.*

Proposition 2.3 (positivity property). *Let Assumption* A1 *be satisfied and let $u(x_0,t_0) > 0$ for some $x_0 \in \mathbb{R}$ and $t_0 > 0$ (respectively, $t_0 \geq 0$ if $u_0 \in C(\mathbb{R})$). Then*

$$u(x_0,t) > 0 \quad \textit{for} \quad t > t_0.$$

Proposition 2.4 (decay rate). *Let assumption* A1 *be satisfied and let $u_0 \in L^1(\mathbb{R})$. Then there exists a constant $C > 0$ which depends only on m, λ, and Λ such that*

$$u(x,t) \leq C\|u_0\|_{L^1(\mathbb{R})} t^{-1/(m+1)} \quad \textit{for} \quad (x,t) \in Q.$$

Propositions 2.2, 2.3, 2.4 are proved for smooth $A(x,t)$ in [9],[6] and [4],[15] correspondingly.

Proposition 2.5 (concentration of mass [7, Theorems 3.1 and 6.1]). *Let Assumption* A1 *be satisfied, let $u_0 \not\equiv 0$, $u_0 \in L^1(\mathbb{R})$, and let $M > 0$ denote*

$$M = \int_{\mathbb{R}} u_0(x)dx.$$

Let $v(x,t) = \int_{-\infty}^{x} u(s,t)ds$. Then for any $\epsilon > 0$ small enough there exist $a > 0$ and $b > 0$ such that

$$\begin{aligned} &\textit{if} \quad x \geq \eta t^{\frac{1}{m+1}} + a, \quad \textit{then} \quad v \geq M - \epsilon; \\ &\textit{if} \quad x \leq -\eta t^{\frac{1}{m+1}} - b, \quad \textit{then} \quad v \leq \epsilon, \end{aligned}$$

where η is a constant which does not depend on ϵ.

PROPOSITION 2.6 (expanding of mass [7, Theorems 3.2 and 6.1]). *Let Assumption* A1 *be satisfied and let u_0, M, and v be the same as in Proposition* 2.5. *For any $\epsilon > 0$ small enough there exist $a > 0$ and $b > 0$ such that*

$$\begin{aligned}&\text{if}\quad x \le \eta_1 t^{\frac{1}{m+1}} - b, \quad \text{then}\quad v \le M - \epsilon;\\ &\text{if}\quad x \ge -\eta_1 t^{\frac{1}{m+1}} + a, \quad \text{then}\quad v \ge \epsilon,\end{aligned}$$

where η_1 is a constant which does not depend on ϵ.

PROPOSITION 2.7 (lower bound in the set containing most of the mass [7, Theorems 4.1 and 6.2]). *Let Assumption* A1 *be satisfied and let u_0, M, and v be the same as in Proposition* 2.5. *For any $\epsilon > 0$ small enough there exist c_0 and T such that*

$$\text{if } t > T \quad \text{and} \quad \epsilon \le v(x,t) \le M - \epsilon,$$

then

$$u(x,t) \ge \frac{c_0}{t^{\frac{1}{m+1}}}.$$

PROPOSITION 2.8 (large time behavior [7, Theorems 5.2 and 6.3]). *Suppose that $u_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, $u_0 \not\equiv 0$ in $\mathbb{R}$, and that there exists $a \in C(0,M)$ such that*

$$A(x,t) - a\Big(\int_{-\infty}^{x} u(s,t)ds\Big) \to 0 \quad \text{as} \quad t \to \infty, \quad \text{uniformly with}$$
$$\text{respect to } x \text{ satisfying } \epsilon < \int_{-\infty}^{x} u(s,t)ds < M - \epsilon\ .$$

Then there exist positive constants η_0 and η_1 and a function $\overline{R} \in C(\mathbb{R})$ such that

(i) $\overline{R} \in C^2((-\eta_0,\eta_1))$, $\overline{R} > 0$ *in* $(-\eta_0,\eta_1)$, *and* $\overline{R} = 0$ *in* $\mathbb{R}\backslash(-\eta_0,\eta_1)$;

(ii) *for any $C > 0$*

$$t^{1/(m+1)}u(x,t) - \overline{R}(xt^{-1/(m+1)}) \to 0 \quad \text{as} \quad t \to \infty$$

uniformly with respect to x satisfying $-Ct^{1/(m+1)} \le x \le Ct^{1/(m+1)}$;

(iii) *the function $\overline{u}$ defined by*

$$\overline{u}(x,t) = t^{-1/(m+1)}\overline{R}(xt^{-1/(m+1)})$$

is a self-similar solution of the equation

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a\Big(\int_{-\infty}^{x} u(s,t)ds\Big)\frac{\partial u^m}{\partial x}\right)$$

satisfying

$$u(x,0) = M\delta(x),$$

where $\delta(x)$ denotes the Dirac measure.

Remark 2.1. Propositions 2.3, 2.5–2.8 are known for the porous medium equation with $A(x,t) \equiv 1$ (see [1], [10]). The main ingredient in the proofs is the comparison with explicit solutions such as self-similar solutions and separable solutions. In Problem (II), however, there are no specific solutions and a different study is required.

3. Lagrangian coordinates and rescaling. We begin this section with noting the two conservation laws

$$\int_{-\infty}^{\infty} \rho(x,t)dx = \int_{-\infty}^{\infty} \rho_0(x)dx \tag{3.1}$$

and

$$\int_{-\infty}^{\infty} \rho(x,t)T(x,t)dx = \int_{-\infty}^{\infty} \rho_0(x)T_0(x)dx. \tag{3.2}$$

These equalities both hold for the approximating sequences; therefore, by Proposition 2.5 (concentration of mass) and Proposition 2.2 (equicontinuity of the sequence ρ_{ϵ_n}), one may pass to the limit.

Now consider system (1.13) where y denotes the mass variable. Setting

$$\tau = \log(t+1) \ ,$$

we consider y and τ as independent variables and introduce the rescaled density

$$\underline{R}(y,\tau) \equiv (t+1)^{1/(a_1+2)}\hat{\rho}(y,t) \ .$$

Defining $\underline{T}(y,\tau) \equiv \hat{T}(y,t)$, we shall denote $\underline{R}(y,\tau)$ and $\underline{T}(y,\tau)$ by $R(y,\tau)$ and $T(y,\tau)$, and we obtain the following equations for R and T:

$$R_\tau = R^2\big(\varphi_1(T)R^{a_1}R_y\big)_y + \frac{1}{a_1+2}R \quad \text{in} \quad (0,M)\times\mathbb{R}^+ \tag{3.3}$$

and

$$T_\tau = e^{\gamma\tau}\big(\varphi_2(T)R^{a_2+2}T_y\big)_y \quad \text{in} \quad (0,M)\times\mathbb{R}^+, \tag{3.4}$$

where we have set

$$\gamma = \frac{a_1-a_2}{a_2+2} \ . \tag{3.5}$$

We shall denote $R(y,0)$ and $T(y,0)$, respectively, by $R_0(y)$ and $T_0(y)$.

Observe that we have assumed that ρ and T are classical solutions of the original equations, but actually this is only known in the set $\mathcal{P}$ [5, Theorem 3.1]. To make the above transformation precise we first perform it for the smooth approximating solution $(\rho_n(x,t), T_n(x,t))$ of Problem I_{ϵ_n}.

Let

$$y = \int_{-L_{\epsilon_n}}^{x} \rho_n(s,t)ds \in [0,M] \quad \text{for} \quad -L_{\epsilon_n} \le x \le L_{\epsilon_n} \quad \text{and} \quad t \ge 0 \ .$$

It follows from (1.3) that if $-L_\epsilon < x < L_\epsilon$ and $t > 0$, then $0 \le y \le M$. The pair of strictly positive functions $R_n(y,\tau) \equiv (t+1)^{1/(a_1+2)}\rho_n(x,t)$ and $T_n(y,\tau) \equiv T_n(x,t)$ with corresponding initial functions R_{0n} and T_{0n} is the unique classical solution of the problem

$$\begin{cases} R_\tau = R^2(\varphi_1(T)R^{a_1}R_y)_y + \frac{1}{a_1+2}R & \text{in } (0,M)\times\mathbb{R}^+, \\ T_\tau = e^{\gamma\tau}(\varphi_2(T)R^{a_2+2}T_y)_y & \text{in } (0,M)\times\mathbb{R}^+, \\ R_y(0,\tau) = R_y(M,\tau) = T_y(0,\tau) = T_y(M,\tau) = 0 & \text{for } \tau > 0, \\ R(y,0) = R_{0n}(y), \ T(y,0) = T_{0n}(y) & \text{for } 0 < y < M. \end{cases}$$

Proposition 3.1. *Let $R(y,\tau)$, $T(y,\tau)$, $R_n(y,\tau)$ and $T_n(y,\tau)$ be defined as above. Then $R \in C((0,M)\times[0,\infty))$ and $T \in C_{\rm loc}(\mathcal{P}')$, where we have set*

$$\mathcal{P}' := \{(y,\tau) \in (0,M)\times[0,\infty) : R(y,\tau) > 0\},$$

$R, T \in C^{2,1}_{\rm loc}(\mathcal{P}')$, R, T satisfy (3.3) and (3.4) in $\mathcal{P}'\backslash\{t=0\}$ and, as $n\to\infty$,

$$\begin{aligned} &R_n \to R \quad in \quad C_{\rm loc}\big((0,M)\times[0,\infty)\big) \cap C^{2,1}_{\rm loc}(\mathcal{P}'\backslash\{t=0\}),\\ &T_n \to T \quad in \quad C_{\rm loc}(\mathcal{P}') \cap C^{2,1}_{\rm loc}(\mathcal{P}'\backslash\{t=0\})\ . \end{aligned}$$

If $\rho_0(x) > 0$ for all $x \in \mathbb{R}$, then $R, T \in C^{2,1}((0,M)\times[0,\infty))$ and

$$R_n \to R \quad and \quad T_n \to T \quad \text{in} \quad C^{2,1}_{\rm loc}((0,M)\times(0,\infty)) \quad as \quad n\to\infty.$$

The proof follows at once from (1.4), (1.5), and the relations $y_x = \rho$, $y_t = \varphi_1(T)\rho^{a_1}\rho_x$ and

$$\lim_{n\to\infty}\Big\| \int_{-\infty}^{x} \rho(s,t)ds - \int_{-L_{\epsilon_n}}^{x} \rho_n(s,t)ds \Big\|_{L^\infty(Q_{\epsilon_n})} = 0. \qquad \square$$

Applying Propositions 2.4, 2.6, and 2.7 to the first equation in system (I) we get the following result.

Proposition 3.2. *There exists a constant $\mathcal{C}$ such that*

$$R(y,\tau) \le \mathcal{C} \quad in \quad [0,M]\times[0,\infty). \tag{3.6}$$

For any $\epsilon \in (0, \frac{1}{2}M)$ there exist constants $\tau_\epsilon \ge 0$ and $\delta_\epsilon > 0$ such that

$$R \ge \delta_\epsilon \quad in \quad (\epsilon, M-\epsilon)\times(\tau_\epsilon,\infty). \tag{3.7}$$

If $\rho_0(y) > 0$ for $0 < y < M$, we may choose $\tau_\epsilon = 0$.

4. Large time behavior of T. In this section we consider the behavior of the temperature $T(y,\tau)$ as $\tau\to\infty$. The main result is the following theorem.

Theorem 4.1. *Let $M > 0$ be defined by (1.3) and*

$$T_\infty := \frac{1}{M}\int_0^M T_0(y)dy.$$

(i) *If $a_1 \ge a_2$, then*

$$T(\cdot,\tau) \to T_\infty \quad in \quad C_{\rm loc}\big((0,M)\big) \quad as \quad \tau\to\infty.$$

(ii) *If $a_1 < a_2$, then there exists a function $\sigma \in C(0,M)$ such that*

$$\begin{aligned} &T(\cdot,\tau) \to \sigma \quad in \quad C_{\rm loc}\big((0,M)\big) \quad as \quad \tau\to\infty,\\ &\nu_0 \le \sigma \le \nu_1 \quad in \quad (0,M), \quad and \quad \frac{1}{M}\int_0^M \sigma(y)dy = T_\infty. \end{aligned} \tag{4.1}$$

There exist ρ_0 and T_0 such that $\sigma \not\equiv T_\infty$ in $(0,M)$.

We observe that we do not expect *uniform* convergence of T in $(0,M)$ as $\tau \to \infty$; for example, if $\rho_0(x)$ is positive for all $x \in \mathbb{R}$ and if $T_0(x) = 2 + \sin x$, we conjecture that $T(y,\tau)$ oscillates wildly near $y = 0$ and $y = M$.

Proof. Let γ be defined by (3.5) and define $\overline{\tau}(\tau)$ for $\tau \geq 0$ by

$$\overline{\tau} = \begin{cases} \int_0^\tau e^{\gamma s} ds = \gamma^{-1}(e^{\gamma\tau} - 1) & \text{if } \gamma \neq 0, \\ \tau & \text{if } \gamma = 0. \end{cases}$$

Observe that

$$\overline{\tau}(\infty) = \infty \quad \text{if} \quad \gamma \geq 0 \quad (\Leftrightarrow a_1 \geq a_2) \tag{4.2}$$

and

$$\overline{\tau}(\infty) = -\frac{1}{\gamma} < \infty \quad \text{if} \quad \gamma < 0 \quad (\Leftrightarrow a_1 < a_2). \tag{4.3}$$

The function $T(y,\overline{\tau}) \equiv T(y,\tau)$ satisfies the equation

$$T_{\overline{\tau}} = \big(\varphi_2(T) R^{a_2+2} T_y\big)_y \quad \text{in} \quad (0,M) \times \big(0, \overline{\tau}(\infty)\big). \tag{4.4}$$

Let $\epsilon > 0$. By (3.6) and (3.7) there exists $\overline{\tau}_\epsilon \in [0, \overline{\tau}(\infty))$ such that T is a classical solution of the uniformly parabolic equation (4.4) in the set $[\frac{1}{2}\epsilon, M - \frac{1}{2}\epsilon] \times [\overline{\tau}_\epsilon, \overline{\tau}(\infty))$. By standard estimates on linear uniformly parabolic equations [11, Chapter 3, Theorem 10.1], for any $\overline{\tau}^*_\epsilon \in (\overline{\tau}_\epsilon, \overline{\tau}(\infty))$ there exist constants $\alpha \in (0,1)$ and C such that

$$\|T\|_{C^{\alpha,\alpha/2}([\epsilon,M-\epsilon]\times[\overline{\tau}^*_\epsilon,\overline{\tau}^*])} \leq C \quad \text{for} \quad \overline{\tau}^* < \overline{\tau}^*_\epsilon < \overline{\tau}(\infty), \tag{4.5}$$

where C does not depend on $\overline{\tau}^*$.

We begin with part (ii): $a_1 < a_2$. By (4.3) and (4.5) there exists $\sigma \in C^\alpha([\epsilon, M-\epsilon])$ such that

$$T(\cdot, \overline{\tau}) \to \sigma \quad \text{in} \quad C([\epsilon, M-\epsilon]) \quad \text{as} \quad \overline{\tau} \to \overline{\tau}(\infty) < \infty.$$

Since $\epsilon > 0$ is arbitrary and σ does not depend on ϵ, we have proved (4.1). Since $\nu_0 \leq T(y,\tau) \leq \nu_1$ in $(0,M) \times \mathbb{R}^+$ and since T satisfies the conservation law (3.2)

$$\int_0^M T(y,\tau) dy = \int_0^M T_0(y) dy = M T_\infty \quad \text{for} \quad \tau > 0, \tag{4.6}$$

the proof of (ii) is complete if we construct an example in which σ is nonconstant in $(0,M)$.

Let $\rho_0, T_0 \in C^1([0,M])$ such that $\rho_0 > 0$, $T_0' \geq 0$, and $T_0' \not\equiv 0$ in $(0,M)$. By Proposition 3.2 the lower bound (3.7) for $R(y,\tau)$ holds with $\tau_\epsilon = 0$ for all $\epsilon > 0$. By Proposition 3.1. the function $T(y,\overline{\tau})$ is a classical solution of (4.4), and its partial derivative $T_y(y,\overline{\tau})$ satisfies the equation

$$(T_y)_{\overline{\tau}} = \big(\varphi_2(T) R^{a_2+2} (T_y)_y\big)_y + \Big(\big(\varphi_2'(T) T_y R^{a_2+2} + \varphi_2(T)(R^{a_2+2})_y\big) T_y\Big)_y .$$

It follows from the approximation of T by T_n that $T_y \geq 0$ (we may assume that $T_{0n}'(y) > 0$). Then for any $\epsilon > 0$ the equation for T_y is uniformly parabolic in $[\frac{1}{2}\epsilon, M - \frac{1}{2}\epsilon] \times [0, \overline{\tau}(\infty))$ and the strong maximum principle implies that

$$T_y(y, \overline{\tau}) > 0 \quad \text{in} \quad (0,M) \times \big(0, \overline{\tau}(\infty)\big).$$

It follows from the estimates for T_y and R_y, which will be proved in Lemmas 6.1 and 6.2 below, that the coefficients of the equation for T_y (considered as a linear equation) are uniformly bounded in $(\epsilon, M-\epsilon)\times(\frac{1}{2}\overline{\tau}_\infty, \overline{\tau}_\infty)$. Hence T_y satisfies Harnack's inequality [2, Theorem 3], and $\sigma' > 0$ in $(0, M)$, which completes the proof of (ii).

It remains to prove part (i).

If $a_1 \geq a_2$, it follows from (4.2) and (4.5) that for any $\epsilon \in (0, \frac{1}{2}M)$ there exist a sequence $\overline{\tau}_n \to \infty$ as $n \to \infty$ and a function $\sigma \in C^{\alpha,\alpha/2}([\epsilon, M-\epsilon]\times[0,1])$ such that

$$T(y, \overline{\tau}_n + \overline{\tau}) \to \sigma(y, \overline{\tau}) \quad \text{in} \quad C\big([\epsilon, M-\epsilon]\times[0,1]\big) \quad \text{as} \quad n \to \infty.$$

Clearly $\nu_0 \leq \sigma \leq \nu_1$ in $[\epsilon, M-\epsilon]\times[0,1]$.

We claim that

$$\int_0^1 \int_\epsilon^{M-\epsilon} T_y^2(y, \overline{\tau}_n + \overline{\tau})dyd\overline{\tau} \to 0 \quad \text{as} \quad n \to \infty. \tag{4.7}$$

Accepting (4.7) for the moment, we find that

$$T_y(y, \overline{\tau}_n + \overline{\tau}) \to 0 \quad \text{in} \quad L^2\big((\epsilon, M-\epsilon)\times(0,1)\big) \quad \text{as} \quad n \to \infty$$

and hence $\sigma_y \in L^2((\epsilon, M-\epsilon)\times(0,1))$ and

$$\sigma_y = 0 \quad \text{a.e. in} \quad (\epsilon, M-\epsilon)\times(0,1).$$

Thus

$$\sigma(y, \overline{\tau}) = g_\epsilon(\overline{\tau}) \quad \text{a.e. in} \quad (\epsilon, M-\epsilon)\times(0,1)$$

for some function $g_\epsilon \in C([0,1])$, and it follows from the conservation law (4.6) that

$$MT_\infty = \int_\epsilon^{M-\epsilon} T(y, \overline{\tau}_n + \overline{\tau})dy + O(\epsilon) \to Mg_\epsilon(\overline{\tau}) + O(\epsilon) \quad \text{as} \quad n \to \infty.$$

Hence $\sigma \equiv T_\infty$ a.e. in $(0, M)\times(0,1)$ and we obtain part (i) of Theorem 4.1.

It remains to prove (4.7). Let $T_n(y, \overline{\tau})$ be the smooth function $T_n(y, \tau)$ in terms of the variable $\overline{\tau}$ instead of τ. Then $T_n(y, \overline{\tau})$ is the solution of

$$\begin{cases} T_{\overline{\tau}} = (\varphi_2(T)R^{a_2+2}T_y)_y & \text{in } (0, M)\times\mathbb{R}^+, \\ T_y(0, \overline{\tau}) = T_y(M, \overline{\tau}) = 0 & \text{for } \overline{\tau} > 0, \\ T(y, 0) = T_{0n}(y) & \text{for } 0 < y < M. \end{cases}$$

Multiplying the equation by T_n and integrating by parts we get that for any $0 \leq \overline{\tau}_1 \leq \overline{\tau}_2$

$$\int_{\overline{\tau}_1}^{\overline{\tau}_2} \int_0^M \varphi_2 R_n^{a_2+2} T_{ny}^2 dyd\overline{\tau} = \frac{1}{2}\int_0^M T_n^2(y, \overline{\tau}_2)dy - \frac{1}{2}\int_0^M T_n^2(y, \overline{\tau}_2)dy.$$

Now let $n \to \infty$. Then, by Proposition 3.1,

$$\int_0^\infty \int_0^M R^{a_2+2} T_y^2 dyd\overline{\tau} < \infty.$$

We now apply Proposition 3.2 and conclude that for any $\epsilon \in (0, \frac{1}{2}M)$ there exists a constant C_ϵ such that

$$\int_0^\infty \int_\epsilon^{M-\epsilon} T_y^2 dyd\overline{\tau} \leq C_\epsilon. \tag{4.8}$$

From (4.8) follows (4.7). $\square$

5. Proof of Theorem 1.2. In this section we prove our main result.

First we have to go back from the (y,τ)-variables to the (x,t)-variables. Since $y_x = \rho$, the transformation $(x,t) \mapsto (y,\tau)$ is invertible as a map from $\mathcal{P}$ to $\mathcal{P}'$. Thus it follows from Theorem 4.1 and Proposition 3.2 that for any $\epsilon \in (0, \frac{1}{2}M)$

$$T(x,t) - \sigma\Big(\int_{-\infty}^{x} \rho(s,t)ds\Big) \to 0 \quad \text{as} \quad t \to \infty, \quad \text{uniformly with}$$

$$\text{respect to} \quad x \quad \text{satisfying} \quad \epsilon < \int_{-\infty}^{x} \rho(s,t)ds < M - \epsilon. \tag{5.1}$$

Hence we may apply Proposition 2.8 with $m = a_1 + 1$, $u = \rho$, $A(x,t) = \frac{1}{m}\varphi_1(T(x,t))$, and $a(v) = \frac{1}{m}\varphi_1(\sigma(v))$ for $0 < v < M$ (observe that $\rho(x,t)$ is constructed as the limit of the solutions of Problem (I_ϵ), the first equation of which is equivalent to the equation in Problem (II_n) with $m = a_1 + 1$, $u_n = \rho_n$, and $A_n(x,t) = \frac{1}{m}\varphi_1(T_n(x,t))$. So there exists $\overline{R}(\eta)$ with the properties listed in Proposition 2.8. In particular (1.6) holds, which, combined with Proposition 2.5, implies that

$$\begin{aligned} &\int_{-\infty}^{x} \rho(s,t)ds - \int_{-\infty}^{xt^{-1/(a_1+2)}} \overline{R}(z)dz \\ &= \int_{-\infty}^{x} \big(\rho(s,t)ds - t^{-1/(a_1+2)}\overline{R}(st^{-1/(a_1+2)})\big)ds \to 0 \quad \text{as} \quad t \to \infty, \end{aligned}$$

uniformly with respect to $x \in \mathbb{R}$. Hence, by Theorem 4.1, for all $\epsilon > 0$

$$T(x,t) - \sigma\Big(\int_{-\infty}^{xt^{-1/(a_1+2)}} \overline{R}(z)dz\Big) \to 0 \quad \text{as} \quad t \to \infty,$$

uniformly with respect to x satisfying $-(\eta_0 - \epsilon)t^{1/(a_1+2)} \le x \le (\eta_1 - \epsilon)t^{1/(a_1+2)}$; so we have obtained (1.7) with

$$\overline{T}(\eta) := \sigma\Big(\int_{-\infty}^{\eta} \overline{R}(z)dz\Big) \quad \text{for} \quad -\eta_0 < \eta < \eta_1. \qquad \square$$

6. Estimates for the spatial derivatives. In this section we prove estimates for R_y and T_y which were used in section 4.

LEMMA 6.1. *Let $\epsilon \in (0, \frac{1}{2}M)$ and let $\tau_{\epsilon/2} \ge 0$ be defined by Proposition 3.2. Then there exists a constant $C_\epsilon > 0$ such that*

$$|R_y| \le C_\epsilon \quad \text{in} \quad [\epsilon, M-\epsilon] \times [\tau_\epsilon + \epsilon, \infty).$$

Proof. Let $\tau_0 \ge \tau_{\epsilon/2}$. Integrating by parts (3.3) for $R_n(y,\tau)$ in the set $(0,M) \times (\tau_0, \tau_0+1)$, we obtain that

$$\begin{aligned} 2\int_{\tau_0}^{\tau_0+1}\int_0^M \varphi_1(T_n)R_n^{a_1+1}R_{ny}^2 dy d\tau &= \int_0^M R_n(y,\tau_0+1)dy \\ &- \int_0^M R_n(y,\tau_0)dy - \frac{1}{a_1+2}\int_{\tau_0}^{\tau_0+1}\int_0^M R_n dy d\tau. \end{aligned} \tag{6.1}$$

By Propositions 3.1 and 3.2 and by (6.1)

$$\begin{aligned}
2\int_{\tau_0}^{\tau_0+1}&\int_{\frac{1}{2}\epsilon}^{M-\frac{1}{2}\epsilon}\varphi_1(T)R^{a_1+1}R_y^2dyd\tau\\
&=2\lim_{n\to\infty}\int_{\tau_0}^{\tau_0+1}\int_{\frac{1}{2}\epsilon}^{M-\frac{1}{2}\epsilon}\varphi_1(T_n)R_n^{a_1+1}R_{ny}^2dyd\tau\\
&\le 2\lim_{n\to\infty}\int_{\tau_0}^{\tau_0+1}\int_0^M\varphi_1(T_n)R_n^{a_1+1}R_{ny}^2dyd\tau\\
&=\lim_{n\to\infty}\Big(\int_0^M R_n(y,\tau_0+1)dy-\int_0^M R_n(y,\tau_0)dy-\frac{1}{a_1+2}\int_{\tau_0}^{\tau_0+1}\int_0^M R_ndyd\tau\Big)\\
&=\int_0^M R(y,\tau_0+1)dy-\int_0^M R(y,\tau_0)dy-\frac{1}{a_1+2}\int_{\tau_0}^{\tau_0+1}\int_0^M Rdyd\tau,
\end{aligned}$$

and hence, by Proposition 3.2,

$$\int_{\tau_0}^{\tau_0+1}\int_{\frac{1}{2}\epsilon}^{M-\frac{1}{2}\epsilon}R_y^2dyd\tau\le C_\epsilon \tag{6.2}$$

for some constant C_ϵ which does not depend on $\tau_0\ge\tau_{\epsilon/2}$.

By (6.2) there exists $y_0\in(\frac{1}{2}\epsilon,M-\frac{1}{2}\epsilon)$ such that

$$\int_{\tau_0}^{\tau_0+1}R_y^2(y_0,\tau)d\tau\le\frac{C_\epsilon}{M-\epsilon}.$$

Dividing (3.3) by $-R^2$ we obtain the equation

$$\Big(\frac{1}{R}\Big)_\tau=\Big(\varphi_1(T)R^{a_1+2}\Big(\frac{1}{R}\Big)_y\Big)_y-\frac{1}{(a_1+2)R}\quad\text{in}\quad\Big(\frac{1}{2}\epsilon,M-\frac{1}{2}\epsilon\Big)\times(\tau_{\epsilon/2},\infty).$$

Setting

$$z(y,\tau)=\int_{y_0}^{y}\frac{1}{R(s,\tau)}ds-\int_{\tau_0}^{\tau}\varphi\big(T(y_0,s)\big)R^{a_1}(y_0,s)R_y(y_0,s)ds$$

for $\frac{1}{2}\epsilon\le y\le M-\frac{1}{2}\epsilon$ and $\tau_0\le\tau\le\tau_0+1$, we have that

$$z_y=\frac{1}{R}\quad\text{in}\quad\Big(\frac{1}{2}\epsilon,M-\frac{1}{2}\epsilon\Big)\times(\tau_0,\tau_0+1),$$

and z satisfies in the set $(\frac{1}{2}\epsilon,M-\frac{1}{2}\epsilon)\times(\tau_0,\tau_0+1)$ the equation

(6.3)
$$z_\tau=\varphi_1(T)R^{a_1+2}z_{yy}-\frac{1}{a_1+2}z-\frac{1}{a_1+2}\int_{\tau_0}^{\tau}\varphi_1\big(T(y_0,s)\big)R^{a_1}(y_0,s)R_y(y_0,s)ds.$$

But $\varphi_1(T)R^{a_1+2}$ is uniformly Hölder continuous, and since for any $\tau_0\le\tau\le\tau+h\le\tau_0+1$

$$\Big|\int_{\tau}^{\tau+h}\varphi_1\big(T(y_0,s)\big)R^{a_1}(y_0,s)R_y(y_0,s)ds\Big|\le K\sqrt{h}\int_{\tau}^{\tau+h}R_y^2(y_0,s)ds\le\frac{KC_\epsilon}{M-\epsilon}\sqrt{h}$$

for some K which does not depend on τ_0, the last term at the right-hand side of (6.3) is also Hölder continuous. Hence it follows from standard local estimates for linear uniformly parabolic equations [11, Chapter 4, Theorem 10.1] that

$$z_{yy} \quad \text{is Hölder continuous in} \quad [\epsilon, M-\epsilon] \times [\epsilon, 1],$$

uniformly with respect to $\tau_0 \geq \tau_\epsilon$. The proof is completed by the observation that

$$R_y = -R^2 \Big(\frac{1}{R} \Big)_y = -R^2 z_{yy} \ . \qquad \square$$

For the gradient T_y we obtain a stronger result than the mere bound on T_y.

LEMMA 6.2. *Let σ be defined by Theorem* 4.1. *Then $\sigma \in C^1_{\text{loc}}((0,M))$ and*

$$T_y(\cdot, \tau) \to \sigma' \quad \textit{in} \quad C_{\text{loc}}((0,M)) \quad \textit{as} \quad \tau \to \infty.$$

Proof. Let $\epsilon > 0$ and let τ and $\overline{\tau}_\epsilon$ be defined as in the proof of Theorem 4.1. Setting

$$w(y,\overline{\tau}) = \varphi_2\big(T(y,\overline{\tau})\big) T_y(y,\overline{\tau}) \quad \text{in} \quad \Big(\frac{1}{2}\epsilon, M - \frac{1}{2}\epsilon\Big) \times \big(\overline{\tau}_\epsilon, \overline{\tau}(\infty)\big),$$

we obtain from (4.4) that w satisfies the equation

$$w_{\overline{\tau}} = \big(\varphi_2(T) T_{\overline{\tau}}\big)_y = \big(\varphi_2 R^{a_2+2} w_y + \varphi_2 (R^{a_2+2})_y w\big)_y$$

in $D_\epsilon = (\frac{1}{2}\epsilon, M - \frac{1}{2}\epsilon) \times (\overline{\tau}_\epsilon, \overline{\tau}(\infty))$. By Proposition 3.1 this linear equation is uniformly parabolic in D_ϵ and, by Lemma 6.1, its coefficients are uniformly bounded. Hence it follows from [11, Chapter 3, Theorem 8.1] that w, and hence also T_y, is bounded in $C^\alpha([\epsilon, M-\epsilon] \times [\overline{\tau}^*_\epsilon, \overline{\tau}^*])$ for some $\alpha \in (0,1)$ and for all $\overline{\tau}_\epsilon < \overline{\tau}^*_\epsilon < \overline{\tau}^* < \overline{\tau}(\infty)$; since the bound does not depend on $\overline{\tau}^*$, the result follows immediately. $\square$

Acknowledgments. Part of this paper was written while the first author visited Tel Aviv University and the second author visited the Istituto per le Applicazioni del Calcolo "Mauro Picone" in Rome. We thank both institutions for their hospitality.

REFERENCES

[1] 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, New York, 1986, pp. 1–46.

[2] D.G. ARONSON AND J. SERRIN, *Local behaviour of solutions of quasilinear parabolic equations*, Arch. Rational Mech. Anal., 25 (1967), pp. 81–123.

[3] G.I. BARENBLATT, *On some unsteady motions of a liquid or a gas in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 67–78.

[4] P. BÉNILAN, *Opérateurs accretifs et semigroupes dans les espaces L^p* ($1 \leq p \leq \infty$), Publications de l'Université de Besancon, 1977.

[5] M. BERTSCH AND S. KAMIN, *A system of degenerate parabolic equations*, SIAM J. Math. Anal., 21 (1990), pp. 905–916.

[6] M. BERTSCH AND S. KAMIN, *A positivity property of solutions of a degenerate parabolic equation with nonconstant coefficients*, Adv. Math. Sci. Appl., 5 (1995), pp. 487–495.

[7] M. BERTSCH AND S. KAMIN, *The porous media equation with nonconstant coefficients*, Adv. Differential Equations, 5 (2000), pp. 269–292.

[8] R. DAL PASSO AND L. GIACOMELLI, *Weak solutions of a strongly coupled degenerate parabolic system*, Adv. Differential Equations, 4 (1999), pp. 617–638.

[9] E. DiBenedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.

[10] S. Kamenomostskaya, *The asymptotic behaviour of the solution of the filtration equation,* Israel J. Math., 14 (1973), pp. 76–87.

[11] O.A. Ladyzhenskaja, V.A. Solonnikov and N.N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[12] P. Rosenau and J.M Hyman, *Analysis of nonlinear mass and energy diffusion*, Phys. Rev. A, 32 (1985), pp. 2370–2373.

[13] P. Rosenau and J.M Hyman, *Plasma diffusion across a magnetic field*, Phys. D, 20 (1986), pp. 444–446.

[14] J.L Vazquez, *Asymptotic behaviour and propagation properties of the one-dimensional flow of gas in a porous medium*, Trans. Amer. Math. Soc., 277 (1983), pp. 507–527.

[15] L. Veron, *Coercivité et propriétés régularisantes des semi-groupes non-linéaires dans des espaces de Banach*, Ann. Fac. Sci. Toulouse Math. (5), 1 (1979), pp. 171–200.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 791–820

THE ENTROPY RATE ADMISSIBILITY CRITERION AND THE ENTROPY CONDITION FOR A PHASE TRANSITION PROBLEM: THE ISOTHERMAL CASE[*]

HARUMI HATTORI[†]

Abstract. In this paper, first, we discuss the relation between the entropy rate admissibility criterion and the entropy condition for a phase transition problem. In the context of the Riemann problem, we consider the compatibility of the above admissibility criteria. Then, combining the two criteria, we study the Riemann problems. We discuss the cases where the initial strains are given in the different phases and where they are given in the same phase.

Key words. phase transition, entropy rate admissibility criterion, entropy condition, hyperbolic-elliptic mixed type system

AMS subject classifications. 35L65, 35M10, 73B99

PII. S0036141098341228

1. Introduction.

We study the relation between the entropy rate admissibility criterion and the entropy condition and then discuss the Riemann problems using the two criteria for a system describing a phase transition problem. The system we discuss is a hyperbolic-elliptic mixed type and given by

$$\begin{aligned} v_t - u_x &= 0, \\ u_t - f(v)_x &= 0, \end{aligned} \tag{1.1}$$

where v, u, and f are strain, velocity, and stress, respectively. We assume that f is a smooth nonmonotone function of v as depicted in Figure 1.1. It is important to note that if f' is nonnegative, the system is hyperbolic and if f' is negative, the system is elliptic. In our case there are two intervals $(0,\alpha]$ and $[\beta,\infty)$ where the system is hyperbolic. They are called the α-phase and β-phase, respectively. We assume that f'' is negative in the α-phase and positive in the β-phase. The interval (α,β) is called the spinodal region and physically unobservable. The horizontal line for which the areas A and B are equal is called the Maxwell stress. The values of v in the α-phase and β-phase at which the Maxwell stress intersect f are denoted by v_α and v_β, respectively. The states $(0,v_\alpha]$ and $[v_\beta,\infty)$ are stable, $(v_\alpha,\alpha]$ and $[\beta,v_\beta)$ are metastable, and (α,β) is unstable. The values γ and δ in the α and β-phases are the values of v at which $f(\gamma)=f(\beta)$ and $f(\delta)=f(\alpha)$, respectively. The Riemann problem is a special initial value problem of (1.1) in which the initial data are given by

$$(v,u)(x,0) = \begin{cases} (v_\ell,u_\ell), & x<0, \\ (v_r,u_r), & x>0, \end{cases} \tag{1.2}$$

where v_ℓ, u_ℓ, v_r, and u_r are constants. We consider the two cases where v_ℓ and v_r are given in the different phases and given in the same phase. We seek a self-similar solution in which constant states are separated by the elementary waves and

[*] Received by the editors June 29, 1998; accepted for publication (in revised form) June 23, 1999; published electronically March 21, 2000. This work was supported by U.S. Army DEPSCoR grant DAAH04-94-G-0246 and NSF grant DMS-9704383.

http://www.siam.org/journals/sima/31-4/34122.html

[†] Department of Mathematics, West Virginia University, Morgantown, WV 26506-6310 (hhattori @wvu.edu).

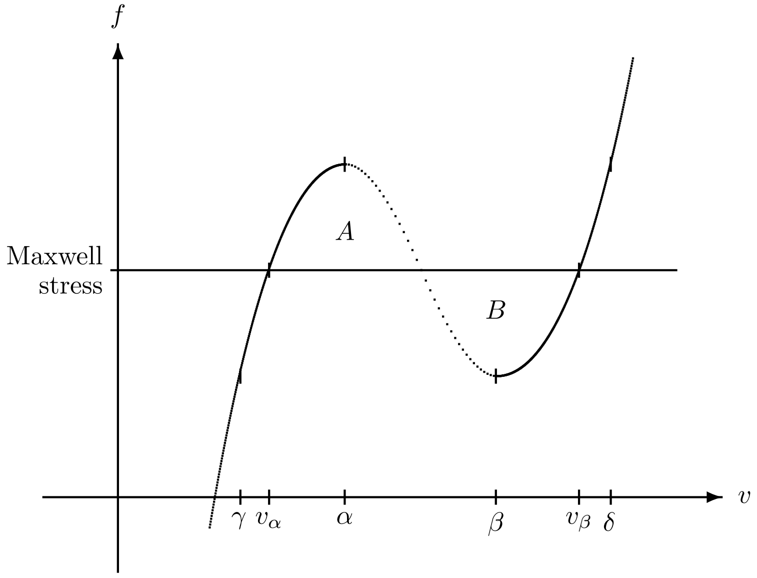


FIG. 1.1.

the phase boundaries. If the solution contains the phase boundaries, we have one or more parameter family of solutions. The above system is an isothermal case. The word "entropy" is not appropriate, yet commonly used; see page 98 in [26]. The nonisothermal system where the thermodynamical effect is taken into account will be discussed in the future.

The purpose of this paper is twofold. First, we study the relation between the entropy rate admissibility criterion and the entropy condition in the context of the Riemann problem. The entropy rate admissibility criterion was proposed by Dafermos [2] for hyperbolic systems. This criterion roughly says that the admissible solution minimizes the rate of entropy decay among the all possible solutions. Both criteria have been used for hyperbolic systems and it is well known that the entropy rate admissibility criterion is more discernible than the entropy condition. On the other hand, the relation between two criteria is not well understood for mixed-type systems. In sections 3 and 4, we study the relation between these criteria for the above mixed-type system. We minimize the entropy rate among the one or more parameter family of solutions if the phase boundaries are involved. It is not clear if the entropy condition is satisfied across phase boundaries. We show, among other things, that there exists a solution minimizing the entropy rate in which the phase boundary violates the entropy condition. We also show that if there are three phase boundaries, at least one phase boundary violates the entropy condition and that there is a solution with three phase boundaries which has a lower entropy rate than the solution with single phase boundary.

Second, the results in section 3 motivate us to consider the Riemann problem using both criteria as the admissibility criterion. In section 4 we propose that the admissible solution to the Riemann problem is the solution that minimizes the entropy rate admissibility criterion among all self-similar solutions satisfying the entropy condition. We call this criterion entropy-entropy rate admissibility criterion. More precisely, we discuss the minimization problems of the following form:

Minimize:	the entropy rate
Subject to:	the entropy condition
	and the characteristic conditions

The specific forms are given in (4.1)–(4.3) in the case of the one-phase boundary problem and in (4.16)–(4.18) in the case of the two-phase boundary problem. In the first case v_ℓ and v_r are specified in the different phases and in the second case v_ℓ and v_r are specified in the same phase. From the results in section 3, there is only one phase boundary in the first case and either two or no phase boundaries in the second case. Imposing both criteria, we obtain a closed and bounded region of parameters in which the entropy rate is minimized. If v_ℓ and v_r are specified in the same phase, there are three cases depending on the relation between the backward and forward wave curves from (v_ℓ, u_ℓ) and (v_r, u_r), respectively. If they do not intersect, two phase boundaries are necessary to solve the Riemann problem. If they intersect at (v_m, u_m), there are two solution configurations. One configuration is the usual one for the hyperbolic system where three constant states (v_ℓ, u_ℓ), (v_m, u_m), and (v_r, u_r) are separated by the backward and forward waves. The other configuration is the one where five constant states are separated by the backward wave, two phase boundaries moving in the opposite directions, and the forward wave. We will show that if v_m is in the metastable state, the solution with two phase boundaries may become admissible. On the other hand, if v_m is in the stable state, the solution with no phase boundary is admissible. Shearer [28] discussed the nonuniqueness of the Riemann problem if v_ℓ and v_r are specified in the same phase. With the combined criterion, we can single out the admissible solution.

Various admissibility criteria have been proposed for hyperbolic systems, for example, the Lax condition, the Liu's condition, the viscosity criterion, the entropy condition, and the entropy rate admissibility criterion; see [18], [22], [2]. The relations among the various admissibility criteria have been discussed extensively in the hyperbolic case. Lax considered the equivalence of Lax shock condition and the entropy criterion [19]. Dafermos discussed this problem for the entropy rate admissibility criterion and the viscosity criterion [2], [3]. It is important to consider whether they can be extended to the mixed-type systems and the relation among these criteria. In this paper we study these problems confining our attention to the criteria based on the entropy.

It is common to model phase transition phenomena in solid from continuum mechanics by using a nonmonotone constitutive relation. The Riemann problem of system (1.1) was discussed in various literature. James [16] initiated the Riemann problem for this type of problem. He proposed the one-parameter family of solutions for the Riemann problem if v_ℓ and v_r are given in the different phases. In the inviscid approaches different admissibility criteria were used to select a physically relevant solution. Abeyaratne and Knowles [1] proposed the kinetic relation for phase boundaries and discussed the Riemann problem using the kinetic relation and the initiation criterion. Hattori [9], [10] used the entropy rate admissibility criterion proposed by Dafermos [2], [3] for hyperbolic systems. Pence [25] and Lin and Pence [21] also used the entropy rate admissibility criterion to discuss phase transition problems. Shearer [27] considered the problem assuming that all the stationary phase boundaries are admissible. Keyfitz [17] discussed the Riemann problem from the point of view of the "hysteresis" approach. As far as the Cauchy problem is concerned, Le Floch [20] has shown the existence of global solutions in BV space if f is trilinear, and Pego and Serre [24] considered the instability of the Glimm scheme. Hattori [12] has shown the existence of weak solutions in the case where f is nonlinear. Another approach is to add the higher spatial derivatives of v and u to smooth out the shock discontinuities and phase boundaries. Slemrod [29], [30] discussed the effects of viscosity and capillarity and proposed the viscosity-capillarity criterion. Shearer [28] considered the issue of nonuniqueness for the Riemann problem using this criterion. Slemrod [31] also discussed the limiting viscosity approach. Fan extended this approach and obtained series of results [4], [5], [6]. He also compared the various admissibility criteria [7]. The results of Fan and Slemrod are summarized in [8]. Hattori and Mischaikow [13] considered the soft loading problem with viscosity and capillarity. Hsiao [15], Hoff and Khodja [14], and Pego [23] considered the role of the viscosity.

This paper consists of 4 sections. In section 2, we discuss preliminary necessary to this paper. In section 3 we discuss the relation between the entropy condition and the entropy rate admissibility criterion. In section 3.1, first we show that if there are three or more phase boundaries for a solution of the Riemann problem, at least one of the phase boundaries violates the entropy condition. Then we discuss the compatibility issue in the case of one phase boundary and a shock wave. We show that there exists an entropy violating phase boundary if the backward wave or forward wave is a shock wave even if the solution satisfies the entropy rate admissibility criterion. In section 3.2, we study the possibility of having three or more phase boundaries in connecting the different phases. Specifically we study the case where there are three phase boundaries connecting different phases. We consider the case where three phase boundaries coalesce and see if a phase boundary is stable against perturbations by three phase boundaries. It turns out that even for a one-phase solution satisfying the entropy rate admissibility criterion, there exists a three-phase boundary solution with lower entropy rate. In section 4, based on the results in section 3, we study the Riemann problems employing the entropy-entropy rate admissibility criterion. In section 4.1, we discuss the case where v_ℓ and v_r are given in the different phases. We show that the entropy condition and the characteristic conditions define the closed and bounded region in the backward (or forward) wave curve where we minimize the entropy rate. In section 4.2, we consider the case where v_ℓ and v_r are given in the same phase. As in section 4.1, we show that there is a closed and bounded region in which the entropy rate is minimized. We show that the results in Theorem 4.7 generalize Fan's result [7] and classify (v_ℓ, u_ℓ) and (v_r, u_r) according to which type of solutions we observe.

2. Preliminary. In this section we summarize the preliminary necessary for this paper.

1. *Elementary waves:* We call the rarefaction wave and the shock wave the elementary waves. The backward wave curve (BWC) $B^r(v_o, u_o)$ is the set of (v, u) connected to (v_o, u_o) on the right by the backward rarefaction or shock wave. It satisfies the following relation:

$$\begin{array}{lll} \text{Rarefaction curve:} & u = u_o + \int_{v_o}^{v} \lambda(w)dw, & \left\{ \begin{array}{ll} v \le v_o & \text{if } f \text{ is convex,} \\ v \ge v_o & \text{if } f \text{ is concave,} \end{array} \right. \\ \text{Shock curve:} & u = u_o - \sigma_b(v_o, v)(v - v_o), & \left\{ \begin{array}{ll} v \ge v_o & \text{if } f \text{ is convex,} \\ v \le v_o & \text{if } f \text{ is concave,} \end{array} \right. \end{array}$$

where $\lambda(w) = \sqrt{f'(w)}$ and $\sigma_b(v_o, v) = -\sqrt{\frac{f(v)-f(v_o)}{v-v_o}}$. The forward wave curve (FWC) $F^r(v_o, u_o)$ is defined in a similar manner:

Rarefaction curve: $u = u_o - \int_{v_o}^{v} \lambda(w)dw,\qquad \begin{cases} v \geq v_o & \text{if } f \text{ is convex,} \\ v \leq v_o & \text{if } f \text{ is concave,} \end{cases}$

Shock curve: $u = u_o - \sigma_f(v_o, v)(v - v_o),\qquad \begin{cases} v \leq v_o & \text{if } f \text{ is convex,} \\ v \geq v_o & \text{if } f \text{ is concave,} \end{cases}$

where $\lambda(w) = \sqrt{f'(w)}$ and $\sigma_f(v_o, v) = \sqrt{\frac{f(v)-f(v_o)}{v-v_o}}$. We define $B^{\ell}(v_o, u_o)$ and $F^{\ell}(v_o, u_o)$ as the sets of (v, u) connected to (v_o, u_o) on the left by the corresponding waves. If the above inequalities are reversed, we obtain the corresponding relations.

2. *Phase boundary:* A phase boundary is the line of discontinuity in the xt-plane across which the phase changes. It satisfies the Rankine–Hugoniot condition. The phase boundary curve $P^r((v_o, u_o))$ (or $P^{\ell}((v_o, u_o))$) is the set of (v, u) connected to (v_o, u_o) on the right (or left) by the phase boundary and satisfies the following relations:

$$u = u_o - \sigma(v_o, v)(v - v_o),$$

where $\sigma(v_o, v) = \pm\sqrt{\frac{f(v)-f(v_o)}{v-v_o}}$ and v_o and v are in the different phases. If the line segment joining $(v_o, f(v_o))$ and $(v, f(v))$ intersect f, the value of v is denoted by v_*.

3. *Admissibility criteria:* The weak solutions for (1.1) are not unique and to choose a physically relevant solution we employ admissibility criteria. There are two criteria that we use in this paper. The entropy rate admissibility criterion is the criterion that was proposed by Dafermos [2], [3]. This criterion roughly says that the rate of entropy (the energy) production is the smallest for the admissible solution. The entropy for (1.1) is given by

$$\eta = \frac{1}{2}u^2 + \int f(v)dv.$$

The rate of decay of the total energy is given by

$$E \equiv D_+\eta = \sum_{\text{jump discontinuities}} \sigma(v_-, v_+)A(v_-, v_+), \tag{2.1}$$

where $\sigma(v_-, v_+)$ is the speed of the jump discontinuity and

$$A(v_-, v_+) = \left[\frac{1}{2}(f(v_-) + f(v_+))(v_+ - v_-) - \int_{v_-}^{v_+} f(w)dw\right].$$

Here v_- and v_+ are the values of v on the left and right of a jump discontinuity. We denote

$$E(v_-, v_+) = \sigma(v_-, v_+)A(v_-, v_+).$$

The entropy rate admissibility criterion postulates that the solution is admissible if it solves (1.1) and minimizes (2.1).

The entropy condition is the criterion imposing that the entropy decreases across discontinuities. This is equivalent to requiring that

$$E(v_-, v_+) = \sigma(v_-, v_+)A(v_-, v_+) \leq 0$$

holds across each discontinuity.

Whenever $A(v_-, v_+) = 0$, the line segment joining $(v_-, f(v_-))$ and $(v_+, f(v_+))$ intersect f. We denote the value of v at which the line segment intersect f by v_e. We assume that v_e is in the elliptic region for all values of v_- and v_+ in the different hyperbolic regions satisfying $A(v_-, v_+) = 0$. We discuss the case where this condition is not satisfied at the end of section 4.

4. *The Riemann problem:* If v_ℓ and v_r are specified in the different phases, the solutions of the Riemann problem consist of the constant states separated by the backward wave, odd numbers of phase boundaries, and the forward wave. If v_ℓ and v_r are specified in the same phase, the solutions of the Riemann problem consist of the constant states separated by the backward wave, even numbers of phase boundaries, and the forward wave. In each case the middle constant states separating the phase boundaries are denoted by (v_i, u_i), $(i = 1, \ldots, n+1)$, where n is the number of phase boundaries and in this case we have an n-parameter family of solutions for the Riemann problem. In Lemma 3.1 it will be shown that there is no backward or forward wave between phase boundaries.

3. Entropy and entropy rate.

3.1. One phase boundary. In this subsection we study the compatibility of the entropy rate admissibility criterion and the entropy condition in the context of the Riemann problem. We mainly discuss the case where v_ℓ and v_r are given in the different phases. We assume without loss of generality that v_ℓ is in the α-phase and v_r is in the β-phase. Specifically, we show that there are cases where we have an entropy violating phase boundary if the forward or backward wave is a shock. Using the convexity of f, we have the following lemma.

Lemma 3.1. *If the two phase boundaries satisfy the entropy condition, there is no backward or forward waves between the two phase boundaries.*

Proof. Denote by (v_1, u_1), (v_2, u_2), and so on the constant states from left to right that are separated by the phase boundaries and waves. Consider the case where v_1 is in the α-phase and the first phase boundary connecting (v_1, u_1) and (v_2, u_2) is a backward phase boundary. The other cases are proved similarly. Since $\sigma(v_1, v_2) \leq 0$, we have $A(v_1, v_2) \geq 0$. Therefore, if there is a wave connecting (v_2, u_2) and (v_3, u_3), the wave speed is larger than $|\sigma(v_1, v_2)|$ and this implies the wave must move forward. Then, the phase boundary connecting (v_3, u_3) and (v_4, u_4) must also move forward. The entropy condition requires that $A(v_3, v_4) = -A(v_4, v_3) < 0$ and the characteristic condition requires that $\lambda_3 \leq \sigma(v_3, v_4)$ or $\sigma(v_2, v_3) \leq \sigma(v_3, v_4)$. It is not difficult to show that it is impossible to satisfy both conditions. For example, if $v_3 < v_2$, we have a forward shock wave connecting (v_2, u_2) and (v_3, u_3). In this case we see that $\lambda_3 < \sigma(v_2, v_3) \leq \sigma(v_3, v_4)$ holds. Then, we can easily show that from the convexity of f, $A(v_3, v_4) < 0$ can not be satisfied; see Figure 3.1. The case where $v_3 > v_2$ can be shown similarly.

Lemma 3.2. *If the two phase boundaries move in the same direction, one of the phase boundaries violates the entropy condition.*

Proof. Consider the case where the two phase boundaries move backward. The case where two phase boundaries move forward can be treated similarly. Suppose the two phase boundaries are separated by the constant states (v_1, u_1), (v_2, u_2), and (v_3, u_3) from left to right. Since $\sigma(v_1, v_2) < 0$, $A(v_1, v_2) \geq 0$ must hold. Now that $\sigma(v_1, v_2) < \sigma(v_2, v_3) < 0$, we have $v_2 < v_3$ if the phase boundary is subsonic at v_1 ($|\sigma(v_1, v_2)| < \lambda(v_1)$), and $v_2 > v_3$ if the phase boundary supersonic at v_1 ($|\sigma(v_1, v_2)| > \lambda(v_1)$); see Figures 3.2 and 3.3. In the first case, $0 \leq A(v_1, v_2) < A(v_3, v_2)$. Since $A(v_3, v_2) = -A(v_2, v_3)$, we see $\sigma(v_2, v_3)A(v_2, v_3) > 0$. Therefore, the second phase

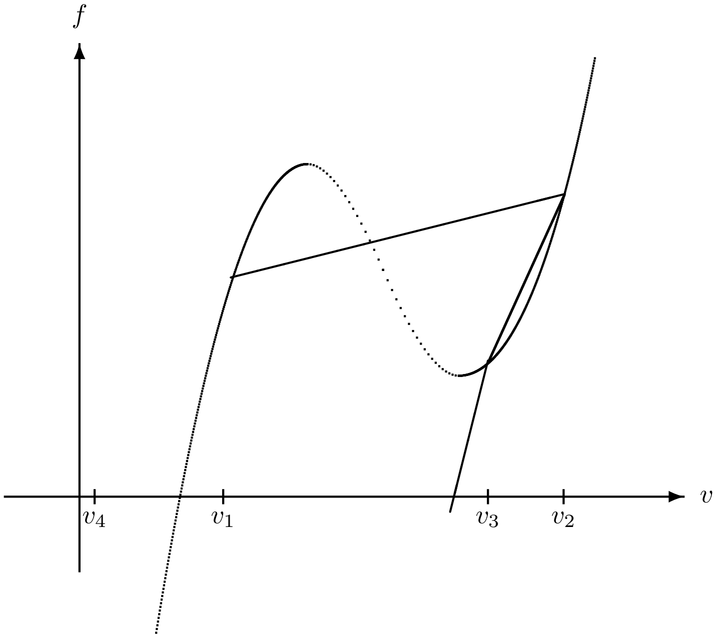


Fig. 3.1.

boundary violates the entropy condition. In the second case, there are two subcases as in Figure 3.3. If we choose $v_3 = v_3'$, then $A(v_2, v_3) < 0$ and $\sigma(v_2, v_3)A(v_2, v_3) > 0$. This violates the entropy condition. If we choose $v_3 = v_3''$, then since we assume that v_e is always in the elliptic region, $A(v_2, v_3) < 0$. This again violates the entropy condition.

Lemma 3.3. *Suppose that v_ℓ is specified in the α-phase and v_r is specified in the β-phase. If there are more than one phase boundary in the resolution of the Riemann problem, at least one of them violates the entropy condition.*

Proof. We consider the case where we have three phase boundaries in the resolution of the Riemann problem. We denote the middle constant states by (v_1, u_1), (v_2, u_2), (v_3, u_3), and (v_4, u_4) from left to right. From the result of Lemma 3.2, the only way that three phase boundaries would be compatible with the entropy condition is if the first phase boundary moves backward, the second phase boundary is stationary, and the third phase boundary moves forward. Since $\sigma(v_1, v_2) < 0$, $A(v_1, v_2) \geq 0$ must hold and $\sigma(v_2, v_3) = 0$, these imply that $A(v_3, v_2) > 0$. Now $v_2 < v_4$, we have $A(v_3, v_4) > 0$ and consequently, $\sigma(v_3, v_4)A(v_3, v_4) > 0$. This is a contradiction.

With the above lemmas as a motivation, we consider the solution to the Riemann problem where v_ℓ and v_r are in the α-phase and β-phase, respectively, and there is only one phase boundary. In what follows we assume that the constant states (v_ℓ, u_ℓ), (v_1, u_1), (v_2, u_2), and (v_r, u_r) are separated by a backward wave, a phase boundary, and a forward wave. We regard u_1 as the independent variable and derive the differential equations for v_1, v_2, and u_2 and the derivatives of the entropy rate. We have four possible cases because there are two possibilities for the backward and forward waves.

(i) Both backward and forward waves are shocks.

(ii) The backward wave is a rarefaction wave and the forward wave is a shock.

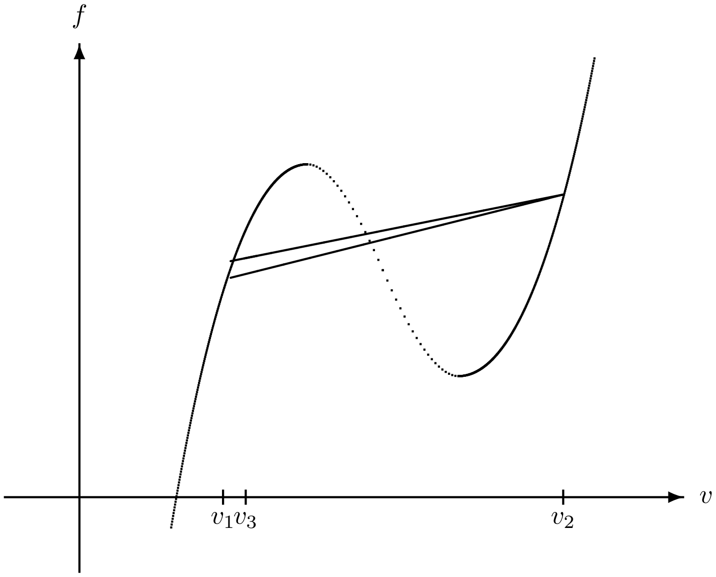


Fig. 3.2.

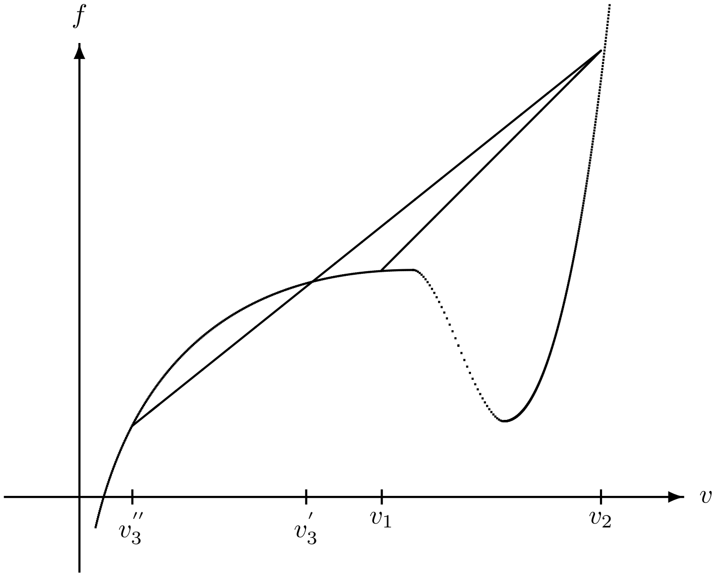


Fig. 3.3.

(iii) The backward wave is a shock and the forward wave is a rarefaction wave.

(iv) Both backward and forward waves are rarefaction waves.

For example, in case (i) we have

$$u_1 = u_\ell - \sigma_b(v_1 - v_\ell), \tag{3.1}$$

$$u_2 = u_1 - \sigma_p(v_2 - v_1), \tag{3.2}$$

$$u_r = u_2 - \sigma_f(v_r - v_2), \tag{3.3}$$

where $\sigma_b = -\sqrt{\frac{f_1 - f_\ell}{v_1 - v_\ell}}$, $\sigma_p = \pm\sqrt{\frac{f_2 - f_1}{v_2 - v_1}}$ (+ for the forward and − for the backward phase boundary), and $\sigma_f = \sqrt{\frac{f_r - f_2}{v_r - v_2}}$. If the backward wave or the forward wave is a rarefaction wave, (3.1) or (3.3) is replaced by

$$u_1 = u_\ell + \int_{v_\ell}^{v_1} \lambda(w)dw \tag{3.4}$$

or

$$u_r = u_2 - \int_{v_2}^{v_r} \lambda(w)dw, \tag{3.5}$$

where $\lambda(w) = \sqrt{f'(w)}$, respectively.

For the backward wave differentiating (3.1) or (3.4), we have

$$\frac{dv_1}{du_1} = -\frac{2\sigma_b}{\lambda_1^2 + \sigma_b^2} \tag{3.6}$$

or

$$\frac{dv_1}{du_1} = \frac{1}{\lambda_1}. \tag{3.7}$$

For the phase boundary differentiating (3.2), we obtain

$$(\lambda_2^2 + \sigma_p^2)\frac{dv_2}{du_1} + 2\sigma_p\frac{du_2}{du_1} = (\lambda_1^2 + \sigma_p^2)\frac{dv_1}{du_1} + 2\sigma_p. \tag{3.8}$$

The derivative of σ_p is expressed as

$$\frac{d\sigma_p}{du_1} = \frac{1 - \frac{du_2}{du_1} - \sigma_p(\frac{dv_2}{du_1} - \frac{dv_1}{du_1})}{(v_2 - v_1)}.$$

For the forward wave differentiating (3.3) or (3.5), we obtain

$$(\lambda_2^2 + \sigma_f^2)\frac{dv_2}{du_1} + 2\sigma_f\frac{du_2}{du_1} = 0 \tag{3.9}$$

or

$$\lambda_2\frac{dv_2}{du_1} + \frac{du_2}{du_1} = 0. \tag{3.10}$$

Combining (3.8) and (3.9), we have

$$\frac{dv_2}{du_1} = \frac{\sigma_f\{2\sigma_p + (\lambda_1^2 + \sigma_p^2)\frac{dv_1}{du_1}\}}{\sigma_f(\lambda_2^2 + \sigma_p^2) - \sigma_p(\lambda_2^2 + \sigma_f^2)}, \tag{3.11}$$

$$\frac{du_2}{du_1} = -\frac{(\lambda_2^2 + \sigma_f^2)\{2\sigma_p + (\lambda_1^2 + \sigma_p^2)\frac{dv_1}{du_1}\}}{2\{\sigma_f(\lambda_2^2 + \sigma_p^2) - \sigma_p(\lambda_2^2 + \sigma_f^2)\}} \tag{3.12}$$

in the case where the forward wave is a shock wave. Combining (3.8) and (3.10), we obtain

$$\frac{dv_2}{du_1} = \frac{\{2\sigma_p + (\lambda_1^2 + \sigma_p^2)\frac{dv_1}{du_1}\}}{(\lambda_2 - \sigma_p)^2}, \tag{3.13}$$

$$\frac{du_2}{du_1} = -\frac{\lambda_2\{2\sigma_p + (\lambda_1^2 + \sigma_p^2)\frac{dv_1}{du_1}\}}{(\lambda_2 - \sigma_p)^2} \tag{3.14}$$

in the case where the forward wave is a rarefaction wave.

The entropy rate is given by

$$E = E_b + E_p + E_f,$$

where

$$E_b = \begin{cases} \sigma_b A(v_\ell, v_1), & v_\ell > v_1, \\ 0, & v_\ell \le v_1, \end{cases}$$

$$E_p = \sigma_p A(v_1, v_2),$$

$$E_f = \begin{cases} \sigma_f A(v_2, v_r), & v_r < v_2, \\ 0, & v_r \ge v_2. \end{cases}$$

The derivatives of entropy rate are given as follows:

$$\frac{dE_b}{du_1} = \begin{cases} \frac{1}{4\sigma_b}(\lambda_1^2 - \sigma_b^2)A_{1\ell}\frac{dv_1}{du_1}, & v_\ell > v_1, \\ 0, & v_\ell \le v_1, \end{cases}$$

$$\frac{dE_p}{du_1} = \frac{1 - \frac{du_2}{du_1} - \sigma_p(\frac{dv_2}{du_1} - \frac{dv_1}{du_1})}{(v_2 - v_1)} A(v_1, v_2) + \sigma_p \frac{dA(v_1, v_2)}{du_1}, \tag{3.15}$$

$$\frac{dE_f}{du_1} = \begin{cases} -\frac{1}{4\sigma_f}(\lambda_2^2 - \sigma_f^2)A_{2r}\frac{dv_2}{du_1}, & v_r < v_2, \\ 0, & v_r \ge v_2, \end{cases}$$

where

$$A_{1\ell} = 3f_1 - f_\ell - \frac{2\int_{v_l}^{v_1} f(w)dw}{v_1 - v_\ell},$$

$$A_{2r} = 3f_2 - f_r - \frac{2\int_{v_2}^{v_r} f(w)dw}{v_r - v_2}.$$

As we can see, if we replace σ_f or σ_b by λ_2 or $-\lambda_1$, respectively, we can obtain the case where we have the rarefaction waves. The following lemma asserts that if v_ℓ and v_r are close to v_α and v_β, respectively, and u_ℓ and u_r are close, the solution of the Riemann problem with $dE/du_1 = 0$ locally minimizes the entropy rate. This was first shown in [11].

LEMMA 3.4. *There exists a neighborhood of $v_\ell = v_\alpha$, $v_r = v_\beta$, $u_\ell = u_c$, and $u_r = u_c$, where u_c is a constant, such that the Riemann problem has a solution which satisfies locally the entropy rate admissibility criterion.*

Proof. Denote dE/du_1 by $F(u_1; v_\ell, v_r, u_\ell, u_r)$. Then, in each case of (i) to (iv) we can easily show that

$$F(u_\alpha; v_\alpha, v_\beta, u_c, u_c) = 0,$$

$$\begin{aligned}\frac{dF}{du_1}(u_\alpha; v_\alpha, v_\beta, u_c, u_c) &= 2\frac{d\sigma_p}{du_1}\frac{dA(v_1,v_2)}{du_1}\\ &= \left(1-\frac{du_2}{du_1}\right)\left(\lambda_1^2\frac{dv_1}{du_1}+\lambda_2^2\frac{dv_2}{du_1}\right)\\ &= 2\frac{\lambda_1(\lambda_1+\lambda_2)}{\lambda_2} > 0,\end{aligned}$$

where u_α is the value of u_1 corresponding to v_α and λ_1 and λ_2 are evaluated at v_α and v_β, respectively. Therefore, the implicit function theorem applies.

The main assumption in Lemma 3.3 was that each discontinuity satisfies the entropy condition. If the middle constant states are chosen so that the entropy rate admissibility criterion is satisfied, one question is whether the phase boundary satisfies the entropy condition.

THEOREM 3.5. *If the forward or backward wave is a shock, there exists a solution to the Riemann problem in which the entropy rate admissibility criterion is locally satisfied but the phase boundary does not satisfy the entropy condition.*

Proof. First consider the case where the both forward and backward waves are rarefaction waves. The derivative of the entropy rate is given by

$$\begin{aligned}\text{(3.16)}\quad \frac{dE}{du_1} &= \frac{1-\frac{du_2}{du_1}-\sigma_p(\frac{dv_2}{du_1}-\frac{dv_1}{du_1})}{(v_2-v_1)}A(v_1,v_2)+\sigma_p\frac{dA(v_1,v_2)}{du_1}\\ &= \frac{\lambda_1+\sigma_p}{\lambda_1(\lambda_2-\sigma_p)}\left\{\frac{(\lambda_1+\lambda_2)A(v_1,v_2)}{(v_2-v_1)}+(v_2-v_1)(\lambda_1\lambda_2+\sigma_p^2)\sigma_p\right\}.\end{aligned}$$

This shows that in order that $\frac{dE}{du_1}=0$, $\sigma_p A(v_1,v_2)\le 0$ must hold. The left-hand side is exactly the entropy rate of the phase boundary and this shows that the entropy condition is satisfied across the phase boundary.

Next, consider the case where at least one of the elementary waves is a shock. As an example we consider the case where f is convex near v_α and v_β and the backward waves is a shock. We construct a solution in which the backward wave is a shock, the phase boundary is on the Maxwell stress, and the forward wave is a rarefaction wave. This construction is possible. First, choose $v_1 = v_\alpha$, $v_2 = v_\beta$, and $u_1 = u_2 = u_c$, where u_c is a constant. Then, choose (v_ℓ, u_ℓ) and (v_r, u_r) on $B^\ell(v_\alpha, u_c)$ and $F^r(v_\beta, u_c)$, respectively, so that we have the backward shock and the forward rarefaction wave. We should choose them so that Lemma 3.4 applies. In this case $v_\ell < v_\alpha$ and $v_r > v_\beta$. The derivative of the entropy rate is given by

$$\frac{dE}{du_1} = \frac{dE_b}{du_1}+\frac{dE_p}{du_1}.$$

At the Maxwell stress $\frac{dE_p}{du_1}=0$, $\frac{d^2E_p}{du_1^2}>0$, and $\frac{dE_b}{du_1}<0$. Therefore, in order that the entropy rate admissibility criterion be satisfied, u_1 must increase. Since $\frac{d^2E_p}{du_1^2}>0$ at the Maxwell stress, $\frac{dE_p}{du_1}>0$ for $u_1>u_\alpha$ and $\frac{dE_p}{du_1}<0$ for $u_1<u_\alpha$. Therefore, the

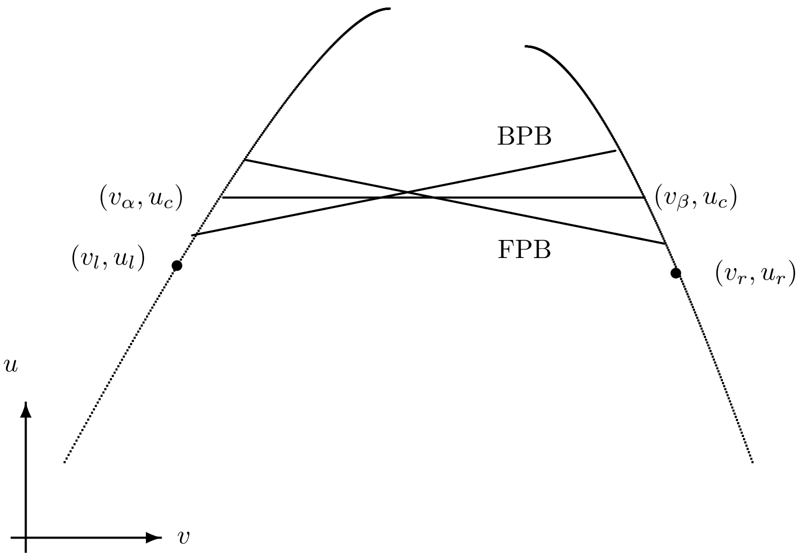


Fig. 3.4.

solution guaranteed by Lemma 3.4 implies that $u_1 > u_\alpha$ should be satisfied in order that the entropy rate admissibility criterion holds. From the differential equations for v_1 and v_2, we see that they increase as we increase u_1. Then from Figure 3.4 we see that we must have the forward phase boundary. Since $\sigma_p > 0$ and $f(v_1)$ and $f(v_2)$ are above the Maxwell line, E_p is positive and it violates the entropy condition.

3.2. Three phase boundaries. In this subsection, we examine the perturbation of a phase boundary against three phase boundaries. It is still interesting to see if there are cases where the solution to the Riemann problem with three phase boundaries has lower entropy rate than the solution with one phase boundary, even though it was shown in section 3.1 that if there are three or more phase boundaries, at least one of them violates the entropy condition. We denote the constant states in the resolution of the Riemann problem by (v_ℓ, u_ℓ), (v_1, u_1), (v_2, u_2), (v_3, u_3), (v_3, u_3), (v_r, u_r). We take u_1, u_2, and u_3 as independent variables and derive the necessary differential equations. We see if the Hessian of E is positive definite when the three phase boundaries coalesce and the coalesced phase boundary gives the local minimum in Lemma 3.4. We obtain the derivatives in the case where both the forward and backward waves are shock waves, since we can take the limit as σ_f approaches λ_4 or σ_b approaches $-\lambda_1$ if we have rarefaction waves. We use $\doteq$ to denote that the quantity is evaluated when the three phase boundaries coalesce.

From the Rankin–Hugoniot conditions we have

$$u_1 = u_\ell - \sigma_b(v_1 - v_\ell), \tag{3.17}$$

$$u_2 = u_1 - \sigma_{p_1}(v_2 - v_1), \tag{3.18}$$

$$u_3 = u_2 - \sigma_{p_2}(v_3 - v_2), \tag{3.19}$$

$$u_4 = u_3 - \sigma_{p_3}(v_4 - v_3), \tag{3.20}$$

$$u_r = u_4 - \sigma_f(v_r - v_4), \tag{3.21}$$

where $\sigma_b < 0$ and $\sigma_f > 0$. If the backward wave or the forward wave is a rarefaction wave, (3.17) or (3.21) is replaced by

$$u_1 = u_\ell + \int_{v_\ell}^{v_1} \lambda(w)dw \tag{3.22}$$

or

$$u_r = u_2 - \int_{v_2}^{v_r} \lambda(w)dw, \tag{3.23}$$

respectively. From (3.17) we have

$$\frac{\partial v_1}{\partial u_1} = -\frac{2\sigma_b}{\lambda_1^2 + \sigma_b^2}, \quad \frac{\partial v_1}{\partial u_2} = 0, \quad \frac{\partial v_1}{\partial u_3} = 0. \tag{3.24}$$

From (3.18) we see

$$\begin{aligned} \frac{\partial v_2}{\partial u_1} &= \frac{2\sigma_{p_1}}{\lambda_2^2 + \sigma_{p_1}^2} + \frac{\lambda_1^2 + \sigma_{p_1}^2}{\lambda_2^2 + \sigma_{p_1}^2}\frac{\partial v_1}{\partial u_1}, \\ \frac{\partial v_2}{\partial u_2} &= -\frac{2\sigma_{p_1}}{\lambda_2^2 + \sigma_{p_1}^2}, \quad \frac{\partial v_2}{\partial u_3} = 0. \end{aligned} \tag{3.25}$$

From (3.19) we see

$$\begin{aligned} \frac{\partial v_3}{\partial u_1} &= \frac{\lambda_2^2 + \sigma_{p_2}^2}{\lambda_3^2 + \sigma_{p_2}^2}\frac{\partial v_2}{\partial u_1}, \\ \frac{\partial v_3}{\partial u_2} &= \frac{2\sigma_{p_2}}{\lambda_3^2 + \sigma_{p_2}^2} + \frac{\lambda_2^2 + \sigma_{p_2}^2}{\lambda_3^2 + \sigma_{p_2}^2}\frac{\partial v_2}{\partial u_2}, \\ \frac{\partial v_3}{\partial u_3} &= -\frac{2\sigma_{p_2}}{\lambda_3^2 + \sigma_{p_2}^2}. \end{aligned} \tag{3.26}$$

From (3.20) and (3.21) we have

$$\begin{aligned} 2\sigma_{p_3}\frac{\partial u_4}{\partial u_1} + (\lambda_4^2 + \sigma_{p_3}^2)\frac{\partial v_4}{\partial u_1} &= (\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_1}, \\ 2\sigma_f\frac{\partial u_4}{\partial u_1} + (\lambda_4^2 + \sigma_f^2)\frac{\partial v_4}{\partial u_1} &= 0, \end{aligned}$$

$$\begin{aligned} \frac{\partial v_4}{\partial u_1} &= \frac{\sigma_f(\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_1}}{\{\sigma_f(\lambda_4^2 + \sigma_{p_3}^2) - \sigma_{p_3}(\lambda_4^2 + \sigma_f^2)\}}, \\ \frac{\partial u_4}{\partial u_1} &= -\frac{(\lambda_4^2 + \sigma_f^2)(\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_1}}{2\{\sigma_f(\lambda_4^2 + \sigma_{p_3}^2) - \sigma_{p_3}(\lambda_4^2 + \sigma_f^2)\}}. \end{aligned} \tag{3.27}$$

In the same way we have

$$\begin{aligned} \frac{\partial v_4}{\partial u_2} &= \frac{\sigma_f(\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_2}}{\{\sigma_f(\lambda_4^2 + \sigma_{p_3}^2) - \sigma_{p_3}(\lambda_4^2 + \sigma_f^2)\}}, \\ \frac{\partial u_4}{\partial u_2} &= -\frac{(\lambda_4^2 + \sigma_f^2)(\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_2}}{2\{\sigma_f(\lambda_4^2 + \sigma_{p_3}^2) - \sigma_{p_3}(\lambda_4^2 + \sigma_f^2)\}}, \end{aligned} \tag{3.28}$$

$$\frac{\partial v_4}{\partial u_3} = \frac{\sigma_f\{2\sigma_{p_3} + (\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_3}\}}{\{\sigma_f(\lambda_4^2 + \sigma_{p_3}^2) - \sigma_{p_3}(\lambda_4^2 + \sigma_f^2)\}}, \tag{3.29}$$

$$\frac{\partial u_4}{\partial u_3} = -\frac{(\lambda_4^2 + \sigma_f^2)\{2\sigma_{p_3} + (\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_3}\}}{2\{\sigma_f(\lambda_4^2 + \sigma_{p_3}^2) - \sigma_{p_3}(\lambda_4^2 + \sigma_f^2)\}}.$$

From the above relations we see that

$$\frac{\partial v_3}{\partial u_2} = \frac{\partial v_4}{\partial u_2} = \frac{\partial u_4}{\partial u_2} = 0 \quad \text{as far as } \sigma_{p_1} = \sigma_{p_2}$$

and

$$\frac{\partial v_4}{\partial u_3} = \frac{\partial u_4}{\partial u_3} = 0 \quad \text{as far as } \sigma_{p_2} = \sigma_{p_3}.$$

Lemma 3.6. *We have the following relations for the second derivatives of u and v:*

$$\frac{\partial^2 v_3}{\partial u_2^2} = \frac{\partial^2 v_4}{\partial u_2^2} = \frac{\partial^2 u_4}{\partial u_2^2} = 0 \quad \textit{as far as } \sigma_{p_1} = \sigma_{p_2} \tag{3.30}$$

and

$$\frac{\partial^2 v_4}{\partial u_3^2} = \frac{\partial^2 u_4}{\partial u_3^2} = 0 \quad \textit{as far as } \sigma_{p_2} = \sigma_{p_3}. \tag{3.31}$$

Proof. For (3.30) we need to show that

$$\frac{\partial}{\partial u_2}\left\{2\sigma_{p_2} + (\lambda_2^2 + \sigma_{p_2}^2)\frac{\partial v_2}{\partial u_2}\right\} = 0 \quad \text{as far as } \sigma_{p_1} = \sigma_{p_2}.$$

It is easy to see that

$$\frac{\partial(\sigma_{p_2} - \sigma_{p_1})}{\partial u_2} = \frac{1 - \sigma_{p_2}(\frac{\partial v_3}{\partial u_2} - \frac{\partial v_2}{\partial u_2})}{(v_3 - v_2)} - \frac{1 + \sigma_{p_1}(\frac{\partial v_2}{\partial u_2} - \frac{\partial v_1}{\partial u_2})}{(v_2 - v_1)}$$

is zero as far as $\sigma_{p_1} = \sigma_{p_2}$. Therefore,

$$2\sigma_{p_2} + (\lambda_2^2 + \sigma_{p_2}^2)\frac{\partial v_2}{\partial u_2} = \frac{2(\lambda_2^2 - \sigma_{p_1}\sigma_{p_2})(\sigma_{p_2} - \sigma_{p_1})}{\lambda_2^2 + \sigma_{p_1}^2}$$

is zero as far as $\sigma_{p_1} = \sigma_{p_2}$. For (3.31) we need to show that

$$\frac{\partial}{\partial u_2}\left\{2\sigma_{p_3} + (\lambda_3^2 + \sigma_{p_3}^2)\frac{\partial v_3}{\partial u_3}\right\} = 0 \quad \text{as far as } \sigma_{p_2} = \sigma_{p_3}.$$

This can be shown in the similar way.

The entropy rate is given by

$$E = E_b + \sigma_{p_1}A(v_1, v_2) + \sigma_{p_2}A(v_2, v_3) + \sigma_{p_3}A(v_3, v_4) + E_f,$$

where

$$E_b = \begin{cases} \sigma_b A(v_\ell, v_1), & v_\ell > v_1, \\ 0, & v_\ell \le v_1, \end{cases}$$

and

$$E_f = \begin{cases} \sigma_f A(v_4, v_r), & v_r > v_4, \\ 0, & v_r \le v_4. \end{cases}$$

As the purpose of this section is to study the perturbation of a single phase boundary against three phase boundaries, in what follows we discuss the case where both backward and forward waves are rarefaction waves.

The first derivatives of E are given by

$$\begin{aligned} \frac{\partial E}{\partial u_1} = {} & \frac{\partial \sigma_{p_1}}{\partial u_1} A(v_1, v_2) + \sigma_{p_1} \frac{\partial A(v_1, v_2)}{\partial u_1} \\ & + \frac{\partial \sigma_{p_2}}{\partial u_1} A(v_2, v_3) + \sigma_{p_2} \frac{\partial A(v_2, v_3)}{\partial u_1} \\ & + \frac{\partial \sigma_{p_3}}{\partial u_1} A(v_3, v_4) + \sigma_{p_3} \frac{\partial A(v_3, v_4)}{\partial u_1}, \end{aligned} \tag{3.32}$$

$$\begin{aligned} \frac{\partial E}{\partial u_2} = {} & \frac{\partial \sigma_{p_1}}{\partial u_2} A(v_1, v_2) + \sigma_{p_1} \frac{\partial A(v_1, v_2)}{\partial u_2} \\ & + \frac{\partial \sigma_{p_2}}{\partial u_2} A(v_2, v_3) + \sigma_{p_2} \frac{\partial A(v_2, v_3)}{\partial u_2} \\ & + \frac{\partial \sigma_{p_3}}{\partial u_2} A(v_3, v_4) + \sigma_{p_3} \frac{\partial A(v_3, v_4)}{\partial u_2}, \end{aligned} \tag{3.33}$$

$$\begin{aligned} \frac{\partial E}{\partial u_3} = {} & \frac{\partial \sigma_{p_2}}{\partial u_3} A(v_2, v_3) + \sigma_{p_2} \frac{\partial A(v_2, v_3)}{\partial u_3} \\ & + \frac{\partial \sigma_{p_3}}{\partial u_3} A(v_3, v_4) + \sigma_{p_3} \frac{\partial A(v_3, v_4)}{\partial u_3}. \end{aligned} \tag{3.34}$$

LEMMA 3.7. *If three phase boundaries coalesce, $\frac{\partial E}{\partial u_2} = \frac{\partial E}{\partial u_3} = 0$. Furthermore, if the coalesced phase boundaries is identical to the one in the previous section, $\frac{\partial E}{\partial u_1} = 0$.*

Proof. It is not difficult to show that $\frac{\partial E}{\partial u_2} = \frac{\partial E}{\partial u_3} = 0$ when three phase boundaries coalesce. For $\frac{\partial E}{\partial u_1}$, we show that it reduces to the same form as $\frac{\partial E_p}{\partial u_1}$ in the previous section when the three phase boundaries coalesce. When the three phase boundaries coalesce, we have

$$\begin{aligned} \frac{\partial E}{\partial u_1} \doteq {} & \frac{1 + \sigma_{p_1} \frac{\partial v_1}{\partial u_1}}{(v_2 - v_1)} A(v_1, v_2) + \frac{-\frac{\partial u_4}{\partial u_1} - \sigma_{p_3} \frac{\partial v_4}{\partial u_1}}{(v_4 - v_3)} A(v_3, v_4) \\ & + \sigma_{p_1} \frac{1}{2} \{ f_1'(v_2 - v_1) - (f_2 - f_1) \} \frac{\partial v_1}{\partial u_1} + \sigma_{p_3} \frac{1}{2} \{ f_4'(v_4 - v_3) - (f_4 - f_3) \} \frac{\partial v_4}{\partial u_1}. \end{aligned}$$

Note that when the three phase boundaries coalesce, $\frac{\partial u_4}{\partial u_1}$ and $\frac{\partial v_4}{\partial u_1}$ reduces to $\frac{\partial u_2}{\partial u_1}$ and $\frac{\partial v_2}{\partial u_1}$ in the previous section. Therefore, it is easy to see that when the three phase boundaries coalesce, $\frac{\partial E}{\partial u_1}$ reduces to the same form as $\frac{\partial E_p}{\partial u_1}$ in the previous section.

We consider the case where the three phase boundaries coalesce and the coalesced phase boundary is identical to the phase boundary that attains the local minimum in section 3.1. We denote such values of v_1 and v_2 satisfying Lemma 3.4 by $\bar{v}_1$

and $\bar{v}_2$, respectively, and $\bar{\sigma}_p$ is the speed of that phase boundary. We assume that $v_1 = v_3 = \bar{v}_1$, $v_2 = v_4 = \bar{v}_2$, $\lambda_1 = \lambda_3 = \bar{\lambda}_1$, $\lambda_2 = \lambda_4 = \bar{\lambda}_2$, and $\sigma_{p_1} = \sigma_{p_2} = \sigma_{p_3} = \bar{\sigma}_p$. In this case $\frac{\partial E}{\partial u_1} = \frac{\partial E}{\partial u_2} = \frac{\partial E}{\partial u_3} = 0$ holds. Note that the phase boundary $\bar{\sigma}_p$ is close to the Maxwell stress. We compute the Hessian matrix H of E when the three phase boundaries coalesce. Denote the components of the Hessian of E by e_{ij} $(i, j = 1, 2, 3)$.

Lemma 3.8. *Both e_{22} and e_{33} are zero when the three phase boundaries coalesce.*

Proof. We consider $\frac{\partial^2 E}{\partial u_3^2}$. $\frac{\partial^2 E}{\partial u_2^2}$ is proved similarly. From Lemma 3.6 we have

$$\begin{aligned}
\frac{\partial^2 E}{\partial u_3^2} \doteq & -\frac{2\frac{\partial \sigma_{p_2}}{\partial u_3}\frac{\partial v_3}{\partial u_3} + \sigma_{p_2}\frac{\partial^2 v_3}{\partial u_3^2}}{(v_3 - v_2)} A(v_2, v_3) + \frac{1 + \sigma_{p_2}\frac{\partial v_3}{\partial u_3}}{(v_2 - v_3)}\{f_3'(v_3 - v_2) - (f_3 - f_2)\}\frac{\partial v_3}{\partial u_3} \\
& + \sigma_{p_2}\left[\frac{1}{2}\{f_3'(v_3 - v_2) - (f_3 - f_2)\}\frac{\partial^2 v_3}{\partial u_3^2} + \frac{1}{2}f_3''(v_3 - v_2)\frac{\partial v_3}{\partial u_3}\frac{\partial v_3}{\partial u_3}\right] \\
& + \frac{2\frac{\partial \sigma_{p_3}}{\partial u_3}\frac{\partial v_3}{\partial u_3} + \sigma_{p_3}\frac{\partial^2 v_3}{\partial u_3^2}}{(v_4 - v_3)} A(v_3, v_4) + \frac{1 + \sigma_{p_3}\frac{\partial v_3}{\partial u_3}}{(v_4 - v_3)}\{f_3'(v_4 - v_3) - (f_4 - f_3)\}\frac{\partial v_3}{\partial u_3} \\
& + \sigma_{p_3}\left[\frac{1}{2}\{f_3'(v_4 - v_3) - (f_4 - f_3)\}\frac{\partial^2 v_3}{\partial u_3^2} + \frac{1}{2}f_3''(v_4 - v_3)\frac{\partial v_3}{\partial u_3}\frac{\partial v_3}{\partial u_3}\right].
\end{aligned}$$

Noting that $\frac{\partial \sigma_{p_2}}{\partial u_3} = \frac{\partial \sigma_{p_3}}{\partial u_3}$ when the three phase boundaries coalesce, we have $\frac{\partial^2 E}{\partial u_3^2} \doteq 0$.

Next, we compute one of the off-diagonal elements e_{23}.

Lemma 3.9. *The e_{23} component of Hessian is not zero in general when the three phase boundaries coalesce.*

Proof. When three phase boundaries coalesce,

$$\begin{aligned}
\frac{\partial^2 E}{\partial u_2 \partial u_3} \doteq & \left\{-\frac{\partial v_2}{\partial u_2} - \frac{\partial v_3}{\partial u_3} - 2\sigma_{p_2}\frac{\partial v_2}{\partial u_2}\frac{\partial v_3}{\partial u_3} - \sigma_{p_2}\frac{\partial^2 v_3}{\partial u_2 \partial u_3}(v_3 - v_2)\right\}\frac{A(v_2, v_3)}{(v_3 - v_2)^2} \\
& - \frac{1 + \sigma_{p_2}\frac{\partial v_2}{\partial u_2}}{(v_2 - v_3)}\frac{1}{2}\{f_3'(v_3 - v_2) - (f_3 - f_2)\}\frac{\partial v_3}{\partial u_3} \\
& + \frac{1 + \sigma_{p_2}\frac{\partial v_3}{\partial u_3}}{(v_2 - v_3)}\frac{1}{2}\{f_2'(v_3 - v_2) - (f_3 - f_2)\}\frac{\partial v_2}{\partial u_2} \\
& + \sigma_{p_2}\frac{1}{2}(f_2' - f_3')\frac{\partial v_2}{\partial u_2}\frac{\partial v_3}{\partial u_3} \\
& - \left\{\frac{\partial^2 u_4}{\partial u_2 \partial u_3} + \sigma_{p_3}\left(\frac{\partial^2 v_4}{\partial u_2 \partial u_3} - \frac{\partial^2 v_3}{\partial u_2 \partial u_3}\right)\right\}\frac{A(v_3, v_4)}{(v_4 - v_3)} \\
& + \sigma_{p_3}\left\{\frac{1}{2}\{f_4'(v_4 - v_3) - (f_4 - f_3)\}\frac{\partial^2 v_4}{\partial u_2 \partial u_3}\right\}.
\end{aligned}$$

From Theorem 3.5, when we have the rarefaction wave, we see

$$A(v_3, v_4) = -(v_4 - v_3)^2\frac{(\lambda_3\lambda_4 + \sigma_{p_3}^2)\sigma_{p_3}}{\lambda_3 + \lambda_4}. \tag{3.35}$$

The lowest order terms are $O(\bar{\sigma}_p)$ and obtained as follows:

$$\begin{aligned}
\frac{\partial^2 E}{\partial u_2 \partial u_3} \doteq & \frac{1}{2}f_3'\frac{\partial v_3}{\partial u_3} - \frac{1}{2}f_2'\frac{\partial v_2}{\partial u_2} \\
& - \frac{\partial^2 u_4}{\partial u_2 \partial u_3}\frac{A(v_3, v_4)}{(v_4 - v_3)} + \sigma_{p_3}\frac{1}{2}f_4'(v_4 - v_3)\frac{\partial^2 v_4}{\partial u_2 \partial u_3} + O(\bar{\sigma}_p^2)
\end{aligned}$$

$$
\begin{aligned}
&\doteq -\frac{\partial^2 u_4}{\partial u_2 \partial u_3}\frac{A(v_3,v_4)}{(v_4-v_3)} + \sigma_{p_3}\frac{1}{2}f_4'(v_4-v_3)\frac{\partial^2 v_4}{\partial u_2 \partial u_3} + O(\bar{\sigma}_p^2)\\
&\doteq \frac{1}{(\lambda_4-\sigma_{p_3})^2}\left(-2\frac{\lambda_3\lambda_4^2\sigma_{p_3}}{\lambda_3+\lambda_4} + \lambda_4^2\sigma_{p_3}\right) + O(\bar{\sigma}_p^2)\\
&\doteq \left(\frac{\bar{\lambda}_2-\bar{\lambda}_1}{\bar{\lambda}_1+\bar{\lambda}_2}\right)\bar{\sigma}_p + O(\bar{\sigma}_p^2).
\end{aligned}
$$

Therefore, there exists a neighborhood of v_α and v_β where $e_{_{23}}$ component of Hessian is not zero if $\bar{\sigma}_p \neq 0$ provided that $\bar{\lambda}_2 \neq \bar{\lambda}_1$ holds.

The above two lemmas show the following.

LEMMA 3.10. *The Hessian of E is in general not positive definite when the three phase boundaries coalesce.*

Proof. From the above lemmas, it is easy to see that the Hessian has the following form when three phase boundaries coalesce:

$$
\begin{pmatrix} e_{11} & e_{12} & e_{13}\\ e_{12} & 0 & e_{23}\\ e_{13} & e_{23} & 0 \end{pmatrix}.
$$

The characteristic equation for the eigenvalues is given by

$$
g(\lambda) \equiv \lambda^3 - e_{11}\lambda^2 - (e_{12}^2+e_{13}^2+e_{23}^2)\lambda - 2e_{12}e_{13}e_{23} + e_{11}e_{23}^2 = 0.
$$

If e_{23} is not zero,

$$
g'(\lambda) = 3\lambda^2 - 2e_{11}\lambda - (e_{12}^2+e_{13}^2+e_{23}^2) = 0
$$

has two real roots with different signs. Since $g(\lambda)=0$ must have three real roots, at least one of the roots of $g(\lambda)=0$ must be negative.

The above lemma implies that if a vector $\vec{u} = (u_1,u_2,u_3)^T$ satisfies

$$
\vec{u}^T H \vec{u} < 0 \tag{3.36}
$$

and

$$
\nabla(\sigma_{p_{i+1}} - \sigma_{p_i}) \cdot \vec{u} \geq 0, \quad i = 1,2, \tag{3.37}
$$

there is a direction of $\vec{u}$ in which the solutions with three phase boundaries have the lower entropy rate than the solution with one phase boundary. First we obtain the direction of $\vec{u}$ which is compatible with the inequalities

$$
\sigma_{p_1} \leq \sigma_{p_2} \leq \sigma_{p_3}.
$$

For this purpose we compute the gradient vectors of the differences when three phase boundaries coalesce. They are given as follows:

$$
\begin{aligned}
&\nabla(\sigma_{p_3} - \sigma_{p_2})\\
&\doteq \left\langle \frac{-\frac{\partial u_4}{\partial u_1} - \sigma_{p_3}(\frac{\partial v_4}{\partial u_1} - \frac{\partial v_3}{\partial u_1})}{(v_4-v_3)} - \frac{\sigma_{p_2}(\frac{\partial v_3}{\partial u_1} - \frac{\partial v_2}{\partial u_1})}{(v_2-v_3)}, \frac{1+\sigma_{p_2}\frac{\partial v_2}{\partial u_2}}{(v_2-v_3)}, 0 \right\rangle,
\end{aligned} \tag{3.38}
$$

$$
\begin{aligned}
&\nabla(\sigma_{p_2} - \sigma_{p_1})\\
&\doteq \left\langle \frac{\sigma_{p_2}(\frac{\partial v_3}{\partial u_1} - \frac{\partial v_2}{\partial u_1})}{(v_2-v_3)} - \frac{1-\sigma_{p_1}(\frac{\partial v_2}{\partial u_1} - \frac{\partial v_1}{\partial u_1})}{(v_2-v_1)}, 0, \frac{1+\sigma_{p_2}\frac{\partial v_3}{\partial u_3}}{(v_2-v_3)} \right\rangle.
\end{aligned} \tag{3.39}
$$

The gradient vector for $(\sigma_{p_3} - \sigma_{p_1})$ can be obtained by adding the above gradient vectors.

In what follows we discuss an example where we can construct a solution to the Riemann problem with three phase boundaries which has the lower entropy rate than the solution in section 3.1. For this purpose we consider the phase boundary for which $\bar{\sigma}_p(\bar{\lambda}_2 - \bar{\lambda}_1) < 0$.

Theorem 3.11. *Consider the solution to the Riemann problem of section* 3.1 *for which*

$$\bar{\sigma}_p(\bar{\lambda}_2 - \bar{\lambda}_1) < 0 \tag{3.40}$$

is satisfied. Suppose that both the forward and backward waves are rarefaction waves. Assume also that $\bar{v}_1$ and $\bar{v}_2$ are close to v_α and v_β, respectively, so that (3.40) *implies that $\frac{\partial^2 E}{\partial u_2 \partial u_3} < 0$. Then, there exists a solution to the same Riemann problem with three phase boundaries which has a lower entropy rate.*

Proof. Since $\frac{\partial^2 E}{\partial u_2 \partial u_3} = \frac{\partial}{\partial u_2}(\frac{\partial E}{\partial u_3}) < 0$, we see that the direction $\vec{u} = \langle 0, 1, 1\rangle$ is compatible with both (3.36) and (3.37). This shows that as we increase u_2, $\frac{\partial E}{\partial u_3}$ decreases. Since $\frac{\partial E}{\partial u_3} = 0$ when the phase boundaries coalesce, $\frac{\partial E}{\partial u_3} < 0$ as we increase u_2. Therefore, if we now increase u_3, then E decreases.

4. Riemann problems.

4.1. One phase boundary. The results in the previous section motivate that we should use both the entropy rate admissibility criterion and the entropy condition as the admissibility criteria. In this subsection using these criteria, we discuss the Riemann problem (1.1) and (1.2) where v_ℓ and v_r are specified in the α-phase and β-phase, respectively. Lemma 3.3 implies that we need to consider only one phase boundary. Since the both backward and forward waves are constructed so that the entropy condition is satisfied across shock discontinuities, we need to impose that $\sigma_p A(v_1, v_2) \leq 0$ be satisfied across the phase boundary. We also require that the speed of the phase boundary in absolute value is less than or equal to that of the backward and forward wave. We apply the entropy rate admissibility criterion among the solutions satisfying the above conditions. Therefore, we have the following optimization problem:

$$\min E \tag{4.1}$$

subject to the entropy condition

$$\sigma_p A(v_1, v_2) \leq 0, \tag{4.2}$$

the characterisitic conditions

$$\sigma_b \text{ or } -\lambda_1 \leq \sigma_p \leq \sigma_f \text{ or } \lambda_2, \tag{4.3}$$

and

$$\begin{aligned} u_1 &= u_l + \begin{cases} -\sigma_b(v_1 - v_\ell), & v_\ell > v_1, \\ \int_{v_\ell}^{v_1} \lambda(w)dw, & v_\ell \leq v_1, \end{cases} \\ u_2 &= u_1 - \sigma_p(v_2 - v_1), \\ u_r &= u_2 - \begin{cases} \sigma_f(v_r - v_2), & v_r < v_2, \\ \int_{v_2}^{v_r} \lambda(w)dw, & v_r \geq v_2, \end{cases} \end{aligned} \tag{4.4}$$

where

$$E = E_b + \sigma_p A(v_1, v_2) + E_f$$

and

$$E_b = \begin{cases} \sigma_b A(v_\ell, v_1), & v_\ell > v_1, \\ 0, & v_\ell \le v_1, \end{cases}$$

$$E_f = \begin{cases} \sigma_f A(v_2, v_r), & v_r < v_2, \\ 0, & v_r \ge v_2. \end{cases}$$

We take one of v_1, u_1, v_2, u_2 as an independent variable. The admissible solution is the solution to the Riemann problem (1.1) and (1.2) satisfying the above minimization problem. We say that a solution is feasible if it satisfies (4.2) and (4.3). We denote this criterion the entropy-entropy rate admissibility criterion. This type of problem was discussed in [25] in the case where there are no shock waves. It should be noted that if the speed of the phase boundary is close to σ_b or σ_f, the line connecting $(v_1, f(v_1))$ and $(v_2, f(v_2))$ will intersect the graph of f three times in the hyperbolic region. It should be interesting to see that this criterion will choose such a solution as an admissible solution.

We construct the admissible solution for given (v_ℓ, u_ℓ) and (v_r, u_r). For this purpose, we find the region of (v_1, u_1) where (4.2) and (4.3) are satisfied for a given forward wave curve $F^\ell(v_r, u_r)$. We call this region the feasible region. We define the curves called the stationary phase boundary curve, the equal area curve, and the equal speed curve. These curves correspond to the equality signs in (4.2) and (4.3). We can also define these curves for a given backward wave curve.

Definition 4.1. *The stationary phase boundary curve (SC) is the curve consisting of the points (v_1, u_1) satisfying*

$$f(v_1) = f(v_2), \quad u_1 = u_2 \tag{4.5}$$

as (v_2, u_2) moves along the forward wave curve. This v_1 satisfies $\gamma \le v_1 \le \alpha$ and exists if $\beta \le v_2 \le \delta$. We can similarly define the stationary phase boundary curve corresponding to the forward wave curve.

Taking u_2 as an independent variable and differentiating $f(v_1) = f(v_2)$ with respect to u_2, we obtain

$$\frac{dv_1}{du_2} = \frac{\lambda_2^2}{\lambda_1^2} \frac{dv_2}{du_2},$$

where

$$\frac{dv_2}{du_2} = \begin{cases} -\frac{1}{\lambda_2}, & v_2 \le v_r, \\ -\frac{2\sigma_f}{\lambda_2^2 + \sigma_f^2}, & v_2 > v_r. \end{cases}$$

We see that this curve has a negative slope in the vu-plane.

Definition 4.2. *The equal area curve (EAC) is the curve consisting of the points (v_1, u_1) satisfying*

$$A(v_1, v_2) = 0, \quad u_2 = u_1 - \sigma_p(v_2 - v_1) \tag{4.6}$$

as (v_2, u_2) *moves along the forward wave curve.*

Differentiating the equations in (4.6) with respect to u_2, we have

$$\frac{dv_1}{du_2} = -\frac{\lambda_2^2 - \sigma_p^2}{\lambda_1^2 - \sigma_p^2}\frac{dv_2}{du_2},$$

$$\frac{du_1}{du_2} = \frac{\{(\lambda_2^2 + \sigma_p^2)(\lambda_1^2 - \sigma_p^2) + (\lambda_2^2 - \sigma_p^2)(\lambda_1^2 + \sigma_p^2)\}\frac{dv_2}{du_2}}{2\sigma_p(\lambda_2^2 - \sigma_p^2)} + 1.$$

Therefore, we obtain

$$\frac{dv_1}{du_1} = -\frac{\sigma_p(\lambda_2^2 - \sigma_p^2)\frac{dv_2}{du_2}}{(\lambda_1^2\lambda_2^2 - \sigma_p^4)\frac{dv_2}{du_2} + \sigma_p(\lambda_1^2 - \sigma_p^2)}.$$

It is easy to see that as long as $-\lambda_1 < \sigma_p < \lambda_2$,

$$\frac{dv_1}{du_1} < (\text{or } >)\ 0 \quad \text{if} \quad \sigma_p > (\text{or } <)\ 0.$$

Suppose (v_β, u_2) is on the forward wave curve. Then, this curve starts from (v_α, u_2).

There are two types of equal speed curves. They are denoted by ESC-I and ESC-II.

Definition 4.3. *The ESC-*I *is the curve consisting of the points* (v_1, u_1) *satisfying*

$$u_2 = u_1 - \sigma_p(v_2 - v_1), \tag{4.7}$$

where

$$\sigma_p = \begin{cases} -\lambda_1, & v_1 \geq v_\ell, \\ \sigma_b, & v_1 < v_\ell, \end{cases} \tag{4.8}$$

as (v_2, u_2) *moves along the forward wave curve. If* (δ, u_2) *is on the forward wave curve, then this curve starts from* (α, u_2), *and if* $v_1 < v_\ell$, *the line segment joining* $(v_2, f(v_2))$ *and* $(v_1, f(v_1))$ *passes through* $(v_\ell, f(v_\ell))$.

Definition 4.4. *The ESC-*II *is the curve consisting of the points* (v_1, u_1) *satisfying*

$$u_2 = u_1 - \sigma_p(v_2 - v_1), \tag{4.9}$$

where

$$\sigma_p = \begin{cases} \lambda_2, & v_2 \leq v_r, \\ \sigma_f, & v_2 > v_r, \end{cases} \tag{4.10}$$

as (v_2, u_2) *moves along the forward wave curve. If* (β, u_2) *is on the forward wave curve, this curve starts from* (γ, u_2), *and if* $v_r < v_2$, *the line segment joining* $(v_2, f(v_2))$ *and* $(v_1, f(v_1))$ *passes through* $(v_r, f(v_r))$.

Remark 4.1. We can define ESC-I and ESC-II for a backward wave curve in a similar manner. In the case of ESC-I, if (v_1, u_1) moves along the backward wave curve, σ_p is equal to the speed of the forward wave curve, and if $v_2 > v_r$, the line segment joining $(v_2, f(v_2))$ and $(v_1, f(v_1))$ passes through $(v_r, f(v_r))$. It should be noted that for $v_r < v_2$ the ESC-I and the forward wave curve coalesce. In the case of ESC-II, if (v_1, u_1) moves along the backward wave curve, σ_p is equal to the speed of the backward wave curve and if $v_1 < v_\ell$, the line segment joining $(v_2, f(v_2))$ and $(v_1, f(v_1))$ passes through $(v_\ell, f(v_\ell))$.

Differentiating (4.7) and (4.8), we obtain

$$\frac{dv_1}{du_1} = -\frac{2\lambda_1(\lambda_2^2 - \lambda_1^2)\frac{dv_2}{du_2}}{\{(\lambda_2^2 + \lambda_1^2)\frac{dv_2}{du_2} - 2\lambda_1\}f_1''(v_2 - v_1) - 2\lambda_1^2(\lambda_2^2 - \lambda_1^2)\frac{dv_2}{du_2}}, \quad v_1 \geq v_l,$$

and

$$\frac{dv_1}{du_1} = \frac{2\sigma_p(v_1 - v_\ell)(\lambda_2^2 - \sigma_p^2)\frac{dv_2}{du_2}}{\{2\sigma_p + (\lambda_2^2 + \sigma_p^2)\frac{dv_2}{du_2}\}(v_2 - v_\ell)(\lambda_1^2 - \sigma_p^2) - (\lambda_1^2 + \sigma_p^2)(v_1 - v_\ell)(\lambda_2^2 - \sigma_p^2)\frac{dv_2}{du_2}},$$
$$v_1 < v_l,$$

as (v_2, u_2) moves along the forward wave curve. Since $f_1'' < 0$, $\lambda_1^2 > \sigma_p^2$, $\lambda_2^2 > \sigma_p^2$, and $\lambda_2 > \lambda_1$ holds if $v_1 \geq v_l$, $\frac{dv_1}{du_1}$ is positive. If $\sigma_p = -\lambda_1$, the slope of the backward phase boundary touches the graph of f. If $\sigma_p = \sigma_b$, the constant state (v_1, u_1) degenerates to a line in the xt-plane, and we interpret that there is no backward wave and that (v_ℓ, u_ℓ) is directly connected to (v_2, u_2). This phase boundary satisfies the entropy condition. It should be noted that for $v_1 < v_l$ the ESC-I and the backward wave curve from $(v_\ell, u_2 + \sigma_p(v_2 - v_1))$ coalesce.

Differentiating (4.9) and (4.10), we obtain

$$\frac{dv_1}{du_1} = -\frac{2\sigma_p}{\lambda_1^2 + \sigma_p^2}$$

as (v_2, u_2) moves along the forward wave curve. Note that $\frac{dv_1}{du_1}$ is negative. If $\sigma_p = \sigma_f$, the constant state (v_2, u_2) degenerates to a line in the xt-plane, and we interpret that there is no forward wave and that (v_r, u_r) is directly connected to (v_1, u_1). This phase boundary also satisfies the entropy condition.

Before discussing the Riemann problems, we state lemmas concerning the relation between the curves defined above.

LEMMA 4.5. *The EAC and SC do not intersect.*

Proof. Consider the case where the EAC and SC for a given forward wave curve intersect at (v_1, u_1). Since both EAC and SC are for the same forward wave curve from (v_r, u_r), for the connection between (v_r, u_r) and (v_1, u_1) through SC, we have

$$u_r - \left\{ \begin{array}{c} \int_{v_r}^{v_2'} \lambda(w)dw \\ \sigma_{f'}(v_2' - v_r) \end{array} \right\} = u_1, \tag{4.11}$$

where v_2' is the value of v on the forward wave curve satisfying $f(v_2') = f(v_1)$ and $\sigma_{f'} = \sqrt{\frac{f(v_2') - f(v_r)}{v_2' - v_r}}$. We have the rarefaction wave if $v_2' \leq v_r$ and the shock wave if $v_2' > v_r$. We also have

$$u_r - \left\{ \begin{array}{c} \int_{v_r}^{v_2} \lambda(w)dw \\ \sigma_f(v_2 - v_r) \end{array} \right\} - \sigma_p(v_1 - v_2) = u_1, \tag{4.12}$$

where v_2 is the value of v on the forward wave curve. We have the rarefaction wave if $v_2 \leq v_r$ and the shock wave if $v_2 > v_r$. The relation (4.12) holds for every connection between (v_r, u_r) and (v_1, u_1) with the forward wave and the phase boundary including

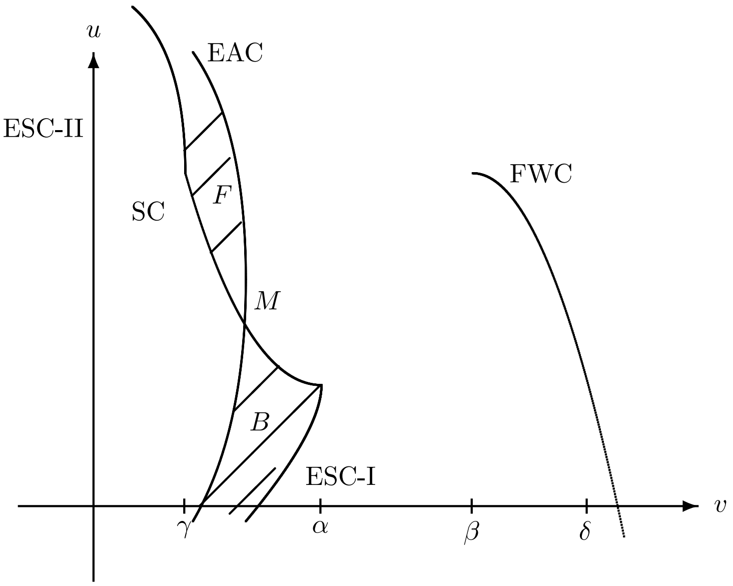


FIG. 4.1.

the one satisfying the equal area condition. From the above two relations, in order for the two curves to meet, we have

$$\left\{ \begin{array}{c} \int_{v_r}^{v_2} \lambda(w)dw \\ \sigma_f(v_2 - v_r) \end{array} \right\} + \sigma_p(v_1 - v_2) = \left\{ \begin{array}{c} \int_{v_r}^{v_2'} \lambda(w)dw \\ \sigma_{f'}(v_2' - v_r) \end{array} \right\}. \tag{4.13}$$

The equality obviously holds if $v_2 = v_2'$. The question is if this holds for other values of v_2. Consider the case where v_r and v_1 are connected by the shock wave. The derivative of the right-hand side with respect to v_2 is given by

$$-\frac{(\sigma_f - \sigma_p)(\lambda_2^2 - \sigma_f\sigma_p)}{2\sigma_f\sigma_p}.$$

Since $|\sigma_p| < \sigma_f < \lambda_2$ holds, the equality in (4.13) holds only when $v_2 = v_2'$. This shows that the EAC and SC do not meet. The case of the rarefaction wave can be shown in a similar way.

THEOREM 4.6. *There exists an absolute minimum for the problems* (4.1), (4.2), (4.3), *and* (4.4). *Furthermore, there exists a neighborhood of $v_\ell = v_\alpha$, $v_r = v_\beta$, $u_\ell = u_c$, and $u_r = u_c$, where u_c is a constant, such that the Riemann problem has a unique solution satisfying the entropy-entropy rate admissibility criterion.*

Proof. Combining the above curves, we obtain the feasible region; see Figure 4.1. In this figure the shaded regions are the feasible regions. Depending on how the backward wave intersects with the shaded region, we obtain three cases.

(a) The backward wave intersects with the region F in Figure 4.1.

(b) The backward wave goes through the point M in Figure 4.1.

(c) The backward wave intersects with the region B in Figure 4.1.

In Figure 4.1, F (or B) stands for the fact that the phase moves forward (or backward) if the backward wave curve intersect this region and the v-coordinate of M is v_α. The

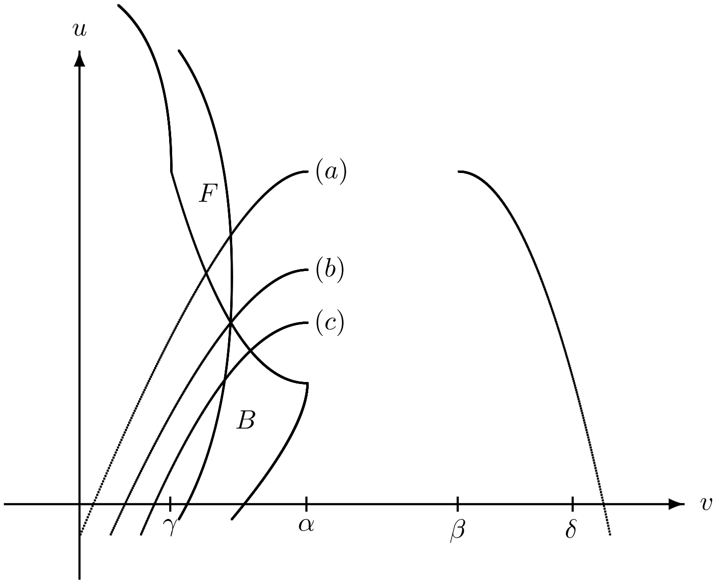


FIG. 4.2.

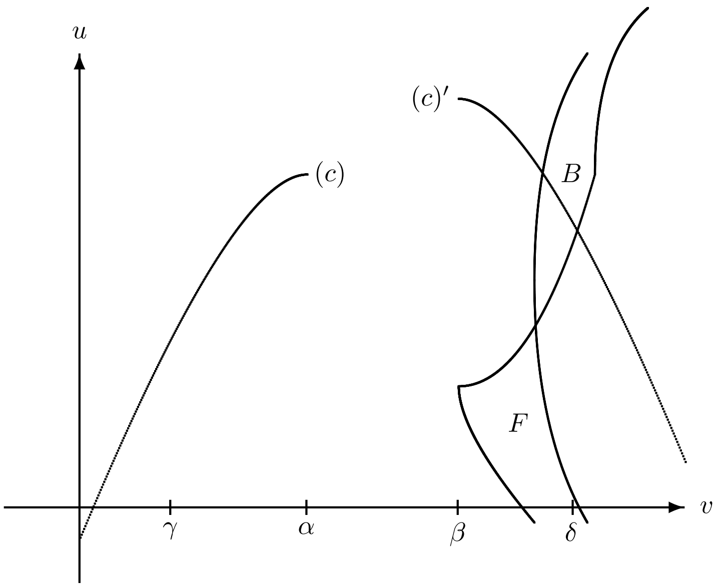


FIG. 4.3.

backward wave curve corresponding to cases (a), (b), and (c) are drawn in Figure 4.2. In case (a) we move (v_1, u_1) along the backward wave curve in the feasible region and find the minimum of the entropy rate. In case (b), only the stationary phase boundary is allowed. In case (c) it is not clear if the backward wave and EAC intersect. To circumvent this difficulty, we change the role of the backward wave curve

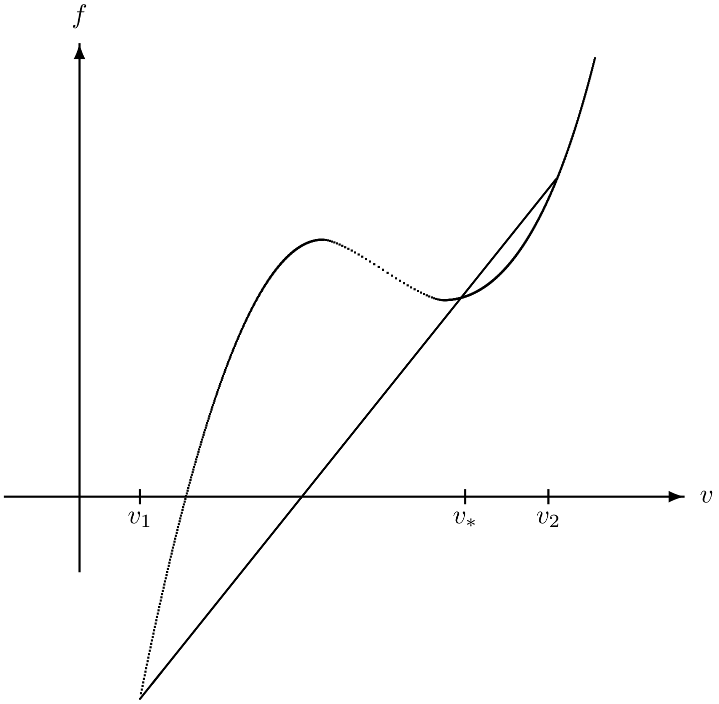


FIG. 4.4.

and the forward wave curve and draw the SC, EAC, and ESC-I, II for the backward wave curve. This is depicted in Figure 4.3. The forward wave curve corresponding to case (c) is given by (c)$'$. It is clear now that we move (v_2, u_2) along the forward wave curve in the feasible region and find the minimum of the entropy rate. The relation between the EAC and ESC-II is not clear. For example, they may intersect. Even if this occurs, the phase boundary connecting (v_1, u_1) and (v_2, u_2) satisfies the entropy condition provided that (v_1, u_1) or (v_2, u_2) is on the ESC-II from the forward wave curve or backward wave curve, respectively. Therefore, an admissible solution always exists. Combining the above result with Lemma 3.4, we see that there exists a neighborhood of $v_\ell = v_\alpha$, $v_r = v_\beta$, $u_\ell = u_c$, and $u_r = u_c$, where u_c is a constant, such that the Riemann problem has a unique solution satisfying the entropy-entropy rate admissibility criterion.

Remark 4.2. If the backward or forward wave curve crosses the ESCs, there are cases where the line segment joining $(v_1, f(v_1))$ and $(v_2, f(v_2))$ intersect with the graph of f at another v in the hyperbolic region. We denote this v by v_*. In this case we may have multiple solutions. One solution consists of v_ℓ, v_1, v_2, and v_r and another solution consists of v_ℓ, v_1, v_*, and v_r. For example, suppose v_* is in the β-phase; see Figure 4.4. For each value of v_r in the β-phase the solution with v_* is not feasible except when $v_* = v_r$ since the characteristic condition (4.3) is not satisfied. If $v_* = v_r$, the constant state v_2 degenerate into a line in the xt-plane. In this case we interpret that v_1 and $v_* = v_r$ are connected by a phase boundary and there is no forward wave.

4.2. Two phase boundaries. In this subsection we construct the solution to the Riemann problem where both v_ℓ and v_r are given in the same phase. We assume that they are given in the α-phase. There are two possibilities. One possibility is that there is no phase boundary and the middle constant state (v_m, u_m) is connected to (v_ℓ, u_ℓ) and (v_r, u_r) by the backward wave and the forward wave, respectively. This is what we observe in the hyperbolic case. This solution will be referred to as the solution with no phase boundary. Another possibility is that the solution has five constant states (v_ℓ, u_ℓ), (v_1, u_1), (v_2, u_2), (v_3, u_3), and (v_r, u_r) which are separated by the backward wave, two phase boundaries, and the forward wave, respectively, from left to right. In this case v_1 and v_3 are in the α-phase and v_2 is in the β-phase. This solution will be referred to as the solution with two phase boundaries. From Lemmas 3.1 and 3.2, there are four possible connections as in section 3. For example, if both the backward and forward waves are shocks, we have

$$u_2 = u_\ell - \sigma_b(v_1 - v_\ell) - \sigma_{p_1}(v_2 - v_1), \tag{4.14}$$

$$u_r = u_2 - \sigma_{p_2}(v_3 - v_2) - \sigma_f(v_r - v_3), \tag{4.15}$$

where $\sigma_b = -\sqrt{\frac{f_1 - f_\ell}{v_1 - v_\ell}}$, $\sigma_{p_1} = -\sqrt{\frac{f_2 - f_1}{v_2 - v_1}}$, $\sigma_{p_2} = \sqrt{\frac{f_3 - f_2}{v_3 - v_2}}$, and $\sigma_f = -\sqrt{\frac{f_3 - f_r}{v_3 - v_r}}$. We regard v_1 and v_3 as functions of v_2 and u_2 and take partial derivatives. Then, we obtain

$$\frac{\partial v_1}{\partial v_2} = \frac{-\sigma_b(\sigma_{p_1}^2 + \lambda_2^2)}{(-\sigma_b + \sigma_{p_1})(\lambda_1^2 - \sigma_b\sigma_{p_1})}, \quad \frac{\partial v_1}{\partial u_2} = \frac{-2\sigma_b\sigma_{p_1}}{(-\sigma_b + \sigma_{p_1})(\lambda_1^2 + \sigma_b\sigma_{p_1})},$$

$$\frac{\partial v_3}{\partial v_2} = \frac{\sigma_f(\sigma_{p_2}^2 + \lambda_2^2)}{(\sigma_f - \sigma_{p_2})(\lambda_3^2 - \sigma_f\sigma_{p_2})}, \quad \frac{\partial v_3}{\partial u_2} = \frac{2\sigma_f\sigma_{p_2})}{(\sigma_f - \sigma_{p_2})(\lambda_3^2 - \sigma_f\sigma_{p_2})}.$$

If the backward or forward wave is a rarefaction wave, we replace σ_b or σ_f with $-\lambda_1$ or λ_2, respectively. Note that there are singularities if $\sigma_{p_1} = \sigma_b$ or $-\lambda_1$ and $\sigma_{p_2} = \sigma_f$ or λ_3.

The optimization problem is given by

$$\min E \tag{4.16}$$

subject to

$$\sigma_{p_1} A(v_1, v_2) \le 0, \quad \sigma_{p_2} A(v_2, v_3) \le 0 \tag{4.17}$$

and

$$\sigma_b \text{ or } -\lambda_1 \le \sigma_{p_1} \le 0 \le \sigma_{p_2} \le \sigma_f \text{ or } \lambda_3, \tag{4.18}$$

and

$$\begin{aligned} u_2 &= u_\ell + \left\{ \begin{array}{c} -\sigma_b(v_1 - v_\ell) \text{ (if } v_\ell > v_1) \\ \int_{v_\ell}^{v_1} \lambda(w)dw \text{ (if } v_\ell \le v_1) \end{array} \right\} - \sigma_{p_1}(v_2 - v_1), \\ u_r &= u_2 - \sigma_{p_2}(v_3 - v_2) - \left\{ \begin{array}{c} \sigma_f(v_r - v_3) \text{ (if } v_r > v_3) \\ \int_{v_3}^{v_r} \lambda(w)dw \text{ (if } v_r \le v_3) \end{array} \right\}, \end{aligned} \tag{4.19}$$

where

$$E = E_b + \sigma_{p_1} A(v_1, v_2) + \sigma_{p_2} A(v_2, v_3) + E_f$$

and

$$E_b = \begin{cases} \sigma_b A(v_\ell, v_1), & v_\ell > v_1, \\ 0, & v_\ell \leq v_1, \end{cases}$$

$$E_f = \begin{cases} \sigma_f A(v_3, v_r), & v_r > v_3, \\ 0, & v_r \leq v_3. \end{cases}$$

We take (v_2, u_2) as independent variables. Lemma 3.2 implies that it is enough to consider the case where two phase boundaries satisfy the condition in (4.18).

As in the previous subsection we draw the backward wave curve and the forward wave curve from (v_ℓ, u_ℓ) and (v_r, u_r), respectively. Denote the intersection point by (v_m, u_m) if it exists. There are three cases depending on the relation between the two curves.

(d) Two curves do not intersect.

(e) Two curves intersect and $v_m > v_\alpha$. Therefore, v_m is in the metastable state.

(f) Two curves intersect and $v_m \leq v_\alpha$. Therefore, v_m is in the stable state.

In case (d) we measure the gap between the BWC and FWC by the difference of u on the BWC and FWC at $v = \alpha$.

Theorem 4.7. *In case* (d) *there exists an absolute minimum for the minimization problem* (4.16), (4.17), (4.18), *and* (4.19). *if the gap between the BWC and FWC is small. In case* (e) *there exists an absolute minimum for the minimization problem which is smaller than the solution with no phase boundary provided that*

$$\frac{\partial E}{\partial v_2} > 0 \tag{4.20}$$

holds at $(v_2, u_2) = (v_s, u_s)$. *On the other hand, in case* (f) *there is no solution with two phase boundaries satisfying the condition* (4.17). *Therefore, the solution with no phase boundary is admissible in this case.*

Proof. In each case we identify the region of (v_2, u_2) where the solution satisfies (4.17) and (4.18). We call this region the feasible region. The construction is similar to the previous subsection. For a given backward wave curve $B^r(v_\ell, u_\ell)$, we draw the curves satisfying the equalities in (4.17) and (4.18). Since $\sigma_{p_1} \leq 0$, we have

$$u_2 = u_1 - \sigma_{p_1}(v_2 - v_1) \geq u_1.$$

This implies that the feasible region lies above the stationary curve of the backward wave curve. Also, we draw the curves satisfying the equalities in (4.17) and (4.18) for a given forward wave curve $F^\ell(v_r, u_r)$. The feasible region lies below the stationary curve of the forward wave curve. The results corresponding to cases (d), (e), and (f) are given in Figures 4.5, 4.6, and 4.7, respectively. In Figure 4.6 the point S at which two stationary phase boundaries meet is denoted by (v_s, u_s). In case (d) we need two phase boundaries to connect (v_ℓ, u_ℓ) and (v_r, u_r). In case (e) the condition (4.20) implies that the solutions with two phase boundaries have lower entropy rates than the solution with no phase boundary provided that (4.20) holds at $v_2 = v_s$, where the constant state (v_2, u_2) degenerates into a vertical line in the xt-plane. Note that

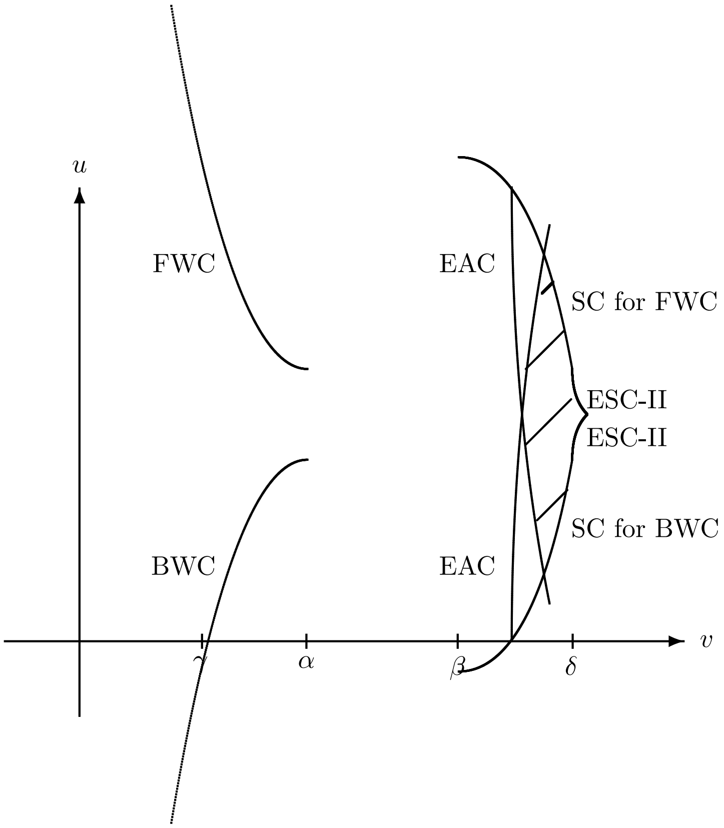

u
FWC
EAC
SC for FWC
ESC-II
ESC-II
SC for BWC
BWC
EAC
v
γ
α
β
δ

Fig. 4.5.

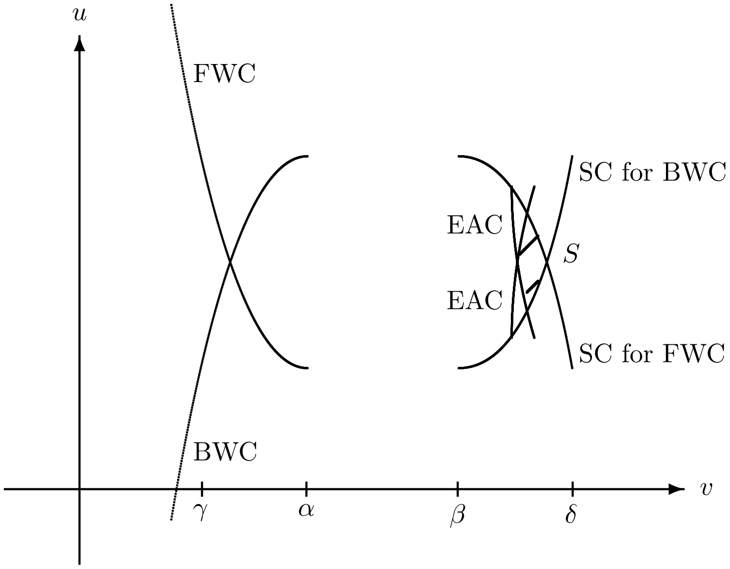

u
FWC
SC for BWC
EAC
S
EAC
SC for FWC
BWC
v
γ
α
β
δ

Fig. 4.6.

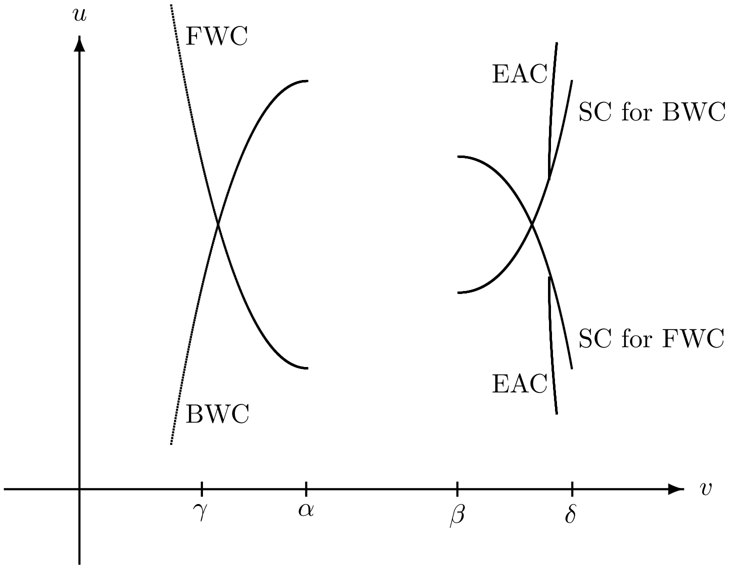


FIG. 4.7.

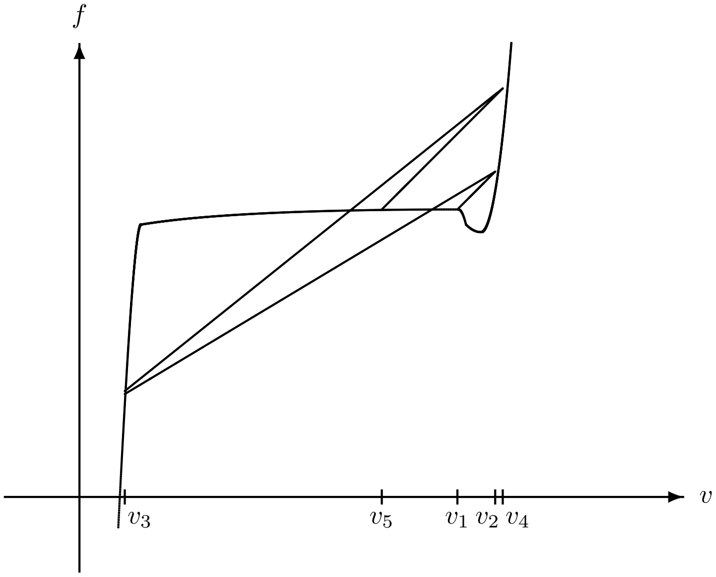


FIG. 4.8.

$\frac{\partial E}{\partial u_2} = 0$ at $(v_2, u_2) = (v_s, u_s)$. In cases (d) and (e) we move (v_2, u_2) in the shaded regions and find the minimum of the entropy rate. In case (f) there is no region where (4.17) and (4.18) are satisfied. In this case we observe the admissible solution with no phase change.

Remark 4.3. In case (d) if the gap between the forward wave and backward wave curves is large, the relation between EAC and ESC-II is not clear. For example, EAC and ESC-II from the backward wave curve or from the forward wave curve may intersect. Even if this occurs, we have an admissible solution as long as two ESC-IIs intersect.

Remark 4.4. In [28] Shearer discussed the nonuniqueness of the Riemann problem. He has observed that when (v_ℓ, u_ℓ) and (v_r, u_r) are connected by the backward rarefaction wave and the forward rarefaction wave, there exists another solution where (v_ℓ, u_ℓ) and (v_r, u_r) are connected by the backward rarefaction wave, the backward phase boundary, the forward phase boundary, and the forward rarefaction wave. He has shown that there are cases where both solutions are admissible if the viscosity-capillarity criterion is employed. Fan [7] compared two types of solutions discussed in Shearer and has shown that the solution with two phase boundaries has the lower entropy rate. The results in Theorem 4.7 generalize Fan's result and classify (v_ℓ, u_ℓ) and (v_r, u_r) according to which type of solutions we observe.

Remark 4.5. It should be noted that Figures 4.1–4.7 are for the case where v_e defined in section 2 is in the spinodal region. If this condition is violated, Lemma 3.2 may not hold. The equal speed curves and the equal area curves may intersect if the gap between the forward and backward waves is large. If this happens, we may observe a situation depicted in Figure 4.8. In this figure, v_1 is connected to v_2 by a backward phase boundary and then v_2 is connected to v_3 by another backward phase boundary. Note that the second phase boundary intersects with the hyperbolic region. If we allow this type of phase boundary, the above two backward phase boundaries satisfy the entropy condition and the solution to the Riemann problem with three or more phase boundaries may become feasible. For example, in Figure 4.8 v_3 can be connected to v_4 by a forward phase boundary and then to v_5 by another forward phase boundary. Therefore, if we identify $v_1 = v_\ell$ and $v_4 = v_r$, and disregard the connection from v_4 to v_5, we see that the solution with three phase boundaries becomes feasible if v_ℓ and v_r are specified in the different phases. If we identify $v_1 = v_\ell$ and $v_5 = v_r$, then we see that the solution with four phase boundaries becomes feasible if v_ℓ and v_r are specified in the same phase.

Acknowledgments. The author is thankful to the referees for their comments through which he was able to improve the paper.

REFERENCES

[1] R. Abeyaratne and J.K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Rational Mech. Anal., 114 (1991), pp. 119–154.

[2] C.M. Dafermos, *The entropy rate admissibility criterion for solutions of hyperbolic conservation laws*, J. Differential Equations, 14 (1973), pp. 202–212.

[3] C.M. Dafermos, *The entropy rate admissibility criterion in thermoelasticity*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend., 8 (1974), pp. 113–119.

[4] H. Fan, *A vanishing viscosity approach on the dynamics of phase transitions in van der Waals fluid*, J. Differential Equations, 103 (1993), pp. 179–204.

[5] H. Fan, *A limiting "viscosity" approach to the Riemann problem for the materials exhibiting change of phase*, Arch. Rational Mech. Anal., 116 (1992), pp. 317–337.

[6] H. Fan, *The uniqueness and stability of the solution of the Riemann problem of a system of conservation laws of mixed type*, Trans. Amer. Math. Soc., 333 (1992), pp. 913–938.

[7] H. Fan, *Global versus local admissibility criteria for dynamic phase boundaries*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 927–944.

[8] H. Fan and M. Slemrod, *The Riemann problem for systems of conservation laws of mixed type*, in Shock Induced Transitions and Phase Structures in General Media, IMA Vol. Math. Appl. 52, Springer, New York, 1993, pp. 61–91.

[9] H. Hattori, *The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion: Isothermal case*, Arch. Rational Mech. Anal., 92 (1986), pp. 247–263.

[10] H. Hattori, *The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion: Non-isothermal case*, J. Differential Equations, 65 (1986), pp. 158–174.

[11] H. Hattori, *An inviscid approach to a phase transition problem*, in Adiabatic Waves in Liquid-Vapor Systems, IUTAM Symposium Gottingen, G.E.A. Meier and P.A. Thompson, eds., Springer-Verlag, Berlin, Heidelberg, 1990, pp. 79-89.

[12] H. Hattori, *The Riemann problem and the existence of weak solutions to a system of mixed-type in dynamic phase transition*, J. Differential Equations, 146 (1998), pp. 287–319.

[13] H. Hattori and K. Mischaikow, *A dynamical system approach to a phase transition problem*, J. Differential Equations, 94 (1991), pp. 340–378.

[14] D. Hoff and M. Khodja, *Stability of coexisting phases for compressible van der Waals fluids*, SIAM J. Appl. Math., 53 (1993), pp. 1–14.

[15] L. Hsiao, *Uniqueness of admissible solutions of the Riemann problem for a system of conservation laws of mixed type*, J. Differential Equations, 86 (1990), pp. 197–233.

[16] R.D. James, *The propagation of phase boundaries in elastic bars*, Arch. Rational Mech. Anal., 73 (1980), pp. 125–158.

[17] B.L. Keyfitz, *The Riemann problem for nonmonotone stress-strain functions: A "hysteresis" approach*, in Nonlinear Systems of Partial Differential Equations in Applied Mathematics, Part I, Lectures in Appl. Math. 23, AMS, Providence, RI, 1986, pp. 379–395.

[18] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[19] P. Lax, *Shock waves and entropy*, in Contributions to Nonlinear Functional Analysis, E. Zarantonello, ed., Academic Press, New York, 1971, pp. 603–634.

[20] P. Le Floch, *Propagating phase boundaries: Formulation of the problem and existence via the Glimm method*, Arch. Rational Mech. Anal., 123 (1993), pp. 153–197.

[21] J. Lin and T.J. Pence, *On the dissipation due to wave ringing in nonelliptic elastic materials*, J. Nonlinear Sci., 3 (1993), pp. 269–305.

[22] T.P. Liu, *The Riemann problem for general systems of conservation laws*, J. Differential Equations, 18 (1975), pp. 218–234.

[23] R.L. Pego, *Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability*, Arch. Rational Mech. Anal., 97 (1987), pp. 353–394.

[24] R.L. Pego and D. Serre, *Instabilities in Glimm's scheme for two systems of mixed type*, SIAM J. Numer. Anal., 25 (1988), pp. 965–988.

[25] T.J. Pence, *On the mechanical dissipation of solutions to the Riemann problem for impact involving a two-phase elastic material*, Arch. Rational Mech. Anal., 117 (1992), pp. 1–52.

[26] M. Renardy and R.C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.

[27] M. Shearer, *The Riemann problem for a class of conservation laws of mixed type*, J. Differential Equations, 46 (1982), pp. 426–443.

[28] M. Shearer, *Nonuniqueness of admissible solutions of Riemann initial value problems for a system of conservation laws of mixed type*, Arch. Rational Mech. Anal., 93 (1986), pp. 45–59.

[29] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[30] M. Slemrod, *Dynamic phase transitions in a van del Waals fluid*, J. Differential Equations, 52 (1984), pp. 1–23.

[31] M. Slemrod, *A limiting "viscosity" approach to the Riemann problem for materials exhibiting change of phase*, Arch. Rational Mech. Anal., 105 (1989), pp. 327–365.

SIAM J. Math. Anal.
Vol. 31, No. 4, pp. 821–835

PARAMETRIC RESONANCE IN WAVE EQUATIONS WITH A TIME-PERIODIC POTENTIAL*

JEFFERY COOPER†

Abstract. We consider the wave equation in three space dimensions perturbed by a time-periodic potential with compact support in space, multiplied by a small parameter, ε. When $\varepsilon = 0$, the scattering theory of Lax and Phillips defines scattering frequencies which describe the decay of solutions in the neighborhood of the support of the potential. For $\varepsilon > 0$, scattering frequencies are defined; they are analogous to Floquet exponents. We show that when the frequency of the time-periodic potential is a multiple of the real part of a scattering frequency σ_0 for the time-independent case, resonance occurs. When ε increases from zero, the scattering frequency σ_0 splits in a symmetric fashion, defining outgoing solutions which decay faster or slower than those of the time-independent problem. An example is given in the case of spherical symmetry of the potential.

Key words. wave equation, scattering, periodic, resonance

AMS subject classifications. 34, 35, 46

PII. S0036141098340703

1. Introduction.

Consider the wave equation with a time-dependent potential

$$u_{tt} - \Delta u + q_0(x)u + \varepsilon p(t)q_1(x)u = 0. \tag{1.1}$$

Here $p(t)$ has period T and ε is a small parameter. We think of this equation as a PDE generalization of Hill's equation

$$u''(t) + q_0 u + \varepsilon p(t)u = 0. \tag{1.2}$$

If the period T of p is a suitable multiple of $2\pi/\sqrt{q_0}$, a resonance occurs, producing an exponentially growing solution of (1.2)(see [9]).

We look for similar behavior for the solutions of (1.1). Our study is motivated by our interest in the behavior of scattering frequencies for (1.1) when q_0 and q_1 have compact support in R^3.

In section 2 we recall some elements of Kato's treatment of analytic perturbations of linear operators [4]. In section 3 we apply this theory to an abstract evolution equation in a complex Hilbert space H,

$$u_t = Au + \varepsilon p(t)Qu. \tag{1.3}$$

We assume that A generates a strongly continuous semigroup of contraction operators $U(t)$ on H and that Q is a bounded operator on H. We assume that A and Q take real vectors into real vectors. For certain values of T, $U(T)$ will have a real eigenvector λ_0. We assume that λ_0 is an isolated point of the spectrum of $U(T)$ with finite multiplicity and no generalized eigenvalues. We show that when p has period T, the eigenvalue λ_0 splits in a symmetric fashion into several branches $\lambda_j(\varepsilon)$ determined by the eigenvectors of A.

[*] Received by the editors June 23, 1998; accepted for publication September 16, 1999; published electronically April 4, 2000.

http://www.siam.org/journals/sima/31-4/34070.html

[†] Department of Mathematics, University of Maryland, College Park, MD 20742 (jec@math.umd.edu).

In section 4, we apply the results of section 3 to (1.1) defined for $(x,t) \in R^3 \times R$ where $q_0(x)$ and $q_1(x)$ have compact support. We do not apply the theory directly to the solutions of (1.1) but rather to a local semigroup $Z(t)$ associated with the solutions of (1.1) with $\varepsilon = 0$. This local semigroup (discussed by Lax and Phillips in [6]) describes the behavior of the solutions in a neighborhood of the support of the scattering potential. In particular $Z(t)$ is compact and has eigenvalues $\exp(i\sigma t)$. The complex numbers σ are called the *scattering frequencies* for (1.1) with $\varepsilon = 0$. The finite energy solutions of (1.1) with $\varepsilon = 0$ decay exponentially in the local energy norm which corresponds to the fact that the scattering frequencies σ satisfy $Im(\sigma) > 0$. Scattering frequencies are also defined for (1.1) when $\varepsilon \neq 0$ (see [1]). In [2], it was shown that the scattering frequencies of (1.1) depend on ε in a continuous fashion. Numerical computations of the scattering frequencies for equations like (1.1) were done in [7] and [8]. These computations showed that the effect of the periodic perturbation may be to force some of the scattering frequencies to the lower half plane, corresponding to exponentially growing solutions of (1.1). In this paper we show that if $\sigma_0 = \nu_0 + i\kappa_0$ is a scattering frequency for (1.1) with $\varepsilon = 0$, with no generalized eigenvectors, and $T = 2\pi/\nu_0$, then σ_0 splits into several branches when $\varepsilon \neq 0$. The directions of the splitting are symmetric with respect to the origin and with respect to the imaginary axis. An example of this splitting is given for spherically symmetric solutions when q_0 and q_1 are real constants. In this case the resonant scattering frequency splits along a vertical line with one branch heading south and one branch heading north. At this time we are still unable to show that as ε increases, one of the resonant scattering frequencies crosses the real axis.

Finally we remark that there is a large literature which treats the Schrödinger equation with a time-periodic potential (for a survey, see the article of Howland [3]). This approach, which uses a quasi energy, did not seem to yield any additional results because of the special nature of the semigroup $Z(t)$. Furthermore our results do not seem to apply to the Schrödinger case because of our hypothesis that the eigenvalue λ_0 be an isolated point of the spectrum of $U(T)$.

2. Analytic perturbation theory.

In this section we recall several results from the theory of analytic perturbations of the spectrum of a bounded operator. The standard reference is Kato [4].

Let $\varepsilon \to L(\varepsilon)$ be a holomorphic family of bounded linear operators on a complex Hilbert space H, defined on a neighborhood of $\varepsilon = 0$. We abbreviate $L(0)$ by writing simply L. $L(\varepsilon)$ may be expanded in a power series, convergent in the operator norm,

$$L(\varepsilon) = L + \varepsilon L_1 + \varepsilon^2 L_2 + \cdots . \tag{2.1}$$

$L_n = L^{(n)}(0)/n!$ are bounded operators on H.

Let $\lambda_0 \in C$ be an isolated point of the spectrum of L which is an eigenvalue of geometric and algebraic multiplicity m. In this case we can apply the finite dimensional theory. λ_0 is a semisimple eigenvalue in the terminology of Kato. The eigenvalue λ_0 may split into several branches $\{\lambda_1(\varepsilon), \ldots, \lambda_s(\varepsilon)\}$, $1 \leq s \leq m$. Let $D = \{|\varepsilon| < \varepsilon_0\}$ and $D_0 = D - \{0\}$.

Theorem 2.1. *Each of the branches $\lambda_j(\varepsilon)$ is differentiable at $\varepsilon = 0$ and holomorphic on D_0 for ε_0 sufficiently small.*

Let Γ be a small circle that encloses λ_0 and $\lambda_j(\varepsilon)$ for $|\varepsilon| \leq \varepsilon_0$, and such that Γ does not meet any other part of the spectrum of $L(\varepsilon)$. Let $R(\zeta, \varepsilon) = (L(\varepsilon) - \zeta)^{-1}$ be the resolvant. Then define

$$P(\varepsilon) = -\frac{1}{2\pi i}\int_\Gamma R(\zeta,\varepsilon)d\zeta. \tag{2.2}$$

$P(\varepsilon)$ is the projection on the total eigenspace corresponding to the eigenvalues $\{\lambda_1(\varepsilon), \ldots, \lambda_s(\varepsilon)\}$.

THEOREM 2.2. *$\varepsilon \to P(\varepsilon)$ is holomorphic on D. The dimension of the range of $P(\varepsilon)$ is constant, equal to m.*

Let E denote the (finite dimensional) eigenspace of λ_0 and let $P = P(0)$ denote the projection onto E. Note that the adjoint of P, P^*, is the projection onto the eigenspace E^* of L^* for the eigenvalue $\bar{\lambda_0}$. It is easy to verify that

$$P^* = -\frac{1}{2\pi i}\int_{\Gamma^*} R^*(\zeta,0)d\zeta, \tag{2.3}$$

where Γ^* is a small circle containing $\bar{\lambda_0}$.

Because λ_0 is a semisimple eigenvalue, the function $\varepsilon \to (L(\varepsilon) - \lambda_0)P(\varepsilon)$ is holomorphic on D and vanishes at $\varepsilon = 0$. Thus it may be expanded in a convergent power series

$$(L(\varepsilon) - \lambda_0)P(\varepsilon) = \sum_1^\infty \varepsilon^n \tilde{L}_n. \tag{2.4}$$

The operators $\tilde{L}_n$ are determined from the L_n as follows. First we expand $R(\zeta,\varepsilon)$ in a power series in ε:

$$R(\zeta,\varepsilon) = R(\zeta) + R_1(\zeta)\varepsilon + R_2(\zeta)\varepsilon^2 + \cdots \tag{2.5}$$

where $R(\zeta) = R(\zeta,0)$ and

$$R_1(\zeta) = -R(\zeta)L_1R(\zeta), \tag{2.6}$$

$$R_2(\zeta) = -R(\zeta)L_2R(\zeta) + R(\zeta)L_1R(\zeta)L_1R(\zeta). \tag{2.7}$$

Now since

$$(L(\varepsilon) - \lambda_0)P(\varepsilon) = -\frac{1}{2\pi i}\int_\Gamma (\zeta - \lambda_0)R(\zeta,\varepsilon)d\zeta, \tag{2.8}$$

we may substitute the expansion (2.5) into (2.8) and use (2.6) and (2.7) to yield the expressions for $\tilde{L_1}$ and $\tilde{L_2}$,

$$\tilde{L_1} = \frac{1}{2\pi i}\int_\Gamma (\zeta - \lambda_0)R(\zeta)L_1R(\zeta)d\zeta, \tag{2.9}$$

$$\tilde{L_2} = \frac{1}{2\pi i}\int_\Gamma (\zeta - \lambda_0)[R(\zeta)L_2R(\zeta) - R(\zeta)L_1R(\zeta)L_1R(\zeta)]d\zeta. \tag{2.10}$$

We can get more explicit expressions for $\tilde{L_1}$ and $\tilde{L_2}$ using the Laurent expansion for $R(\zeta)$ at λ_0:

$$R(\zeta) = -P(\zeta - \lambda_0)^{-1} + \sum_1^\infty S^n(\zeta - \lambda_0)^n, \tag{2.11}$$

where S is the *reduced resolvent* of L. S is defined by $S = (L - \lambda_0)^{-1}(I - P)$ on the range of $I - P$, and $S = 0$ on E. Substitute (2.11) into (2.9) and (2.10) and perform the residue calculation. We obtain

$$\tilde{L}_1 = PL_1P, \tag{2.12}$$

$$\begin{aligned} \tilde{L}_2 &= PL_2P - PL_1PL_1S - PL_1SL_1P - SL_1PL_1P \\ &= PL_2P - \tilde{L}_1L_1S - SL_1\tilde{L}_1 - PL_1SL_1P. \end{aligned} \tag{2.13}$$

THEOREM 2.3. *For each branch $\lambda_j(\varepsilon)$, $\lambda_j'(0)$ is an eigenvalue of $\tilde{L}_1$. Furthermore if $\varphi(\varepsilon)$ is a continuous family of eigenvectors of $L(\varepsilon)$ with eigenvalue $\lambda_j(\varepsilon)$,*

$$L(\varepsilon)\varphi(\varepsilon) = \lambda_j(\varepsilon)\varphi(\varepsilon), \tag{2.14}$$

then $\varphi(0)$ is an eigenvector of $\tilde{L}_1$,

$$\tilde{L}_1\varphi(0) = \lambda_j'(0)\varphi(0). \tag{2.15}$$

Proof. For each ε, let $\varphi(\varepsilon)$ be an eigenvector of $L(\varepsilon)$ with eigenvalue $\lambda_j(\varepsilon)$, that is, (2.14). We may assume that $\|\varphi(\varepsilon)\| = 1$. Then because the unit ball in H is weakly compact, we can extract a subsequence $\varepsilon_k \to 0$ such that $\varphi(\varepsilon_k) \to w$ weakly in H. Since $\|\varphi(\varepsilon_k)\| = 1$ we also have $\varphi(\varepsilon_k) \to w$ strongly in H. We rewrite the left side of (2.4) as

$$\begin{aligned} (L(\varepsilon_k) - \lambda_0)P(\varepsilon_k)\varphi(\varepsilon_k) &= (L(\varepsilon_k) - \lambda_j(\varepsilon_k) + \lambda_j(\varepsilon_k) - \lambda_0)\varphi(\varepsilon_k) \\ &= (\lambda_j(\varepsilon_k) - \lambda_0)\varphi(\varepsilon_k). \end{aligned}$$

Then dividing (2.4) by ε_k we have

$$\frac{\lambda(\varepsilon_k) - \lambda_0}{\varepsilon_k}\varphi(\varepsilon_k) = \tilde{L}_1\varphi(\varepsilon_k) + \sum_{n=2}^{\infty} \varepsilon_k^{n-1}\tilde{L}_n\varphi(\varepsilon_k).$$

We take the limit as $k \to \infty$ and deduce that $\lambda_j'(0)w = \tilde{L}_1 w$. Of course, the same results hold if $\varphi(\varepsilon)$ is any continuous family of eigenvectors satisfying (2.14).

If $\lambda_j'(0)$ is a semisimple eigenvalue of $\tilde{L}_1$, we may apply Theorem 2.1, and deduce that $\lambda_j(\varepsilon)$ is twice differentiable at $\varepsilon = 0$.

THEOREM 2.4. *Assume that $\varphi(\varepsilon)$ is a continuous family of eigenvectors satisfying* (2.14). *Assume that $\tilde{L}_1 = 0$. Then $\lambda_j'(0) = 0$ for each j and $\varphi(0)$ is an eigenvector of $\tilde{L}_2$ with eigenvalue $2\lambda_j''(0)$,*

$$\tilde{L}_2\varphi(0) = 2\lambda_j''(0)\varphi(0). \tag{2.16}$$

Note that in this case, (2.13) becomes

$$\tilde{L}_2 = PL_2P - PL_1SL_1P. \tag{2.17}$$

Proof. Since $\tilde{L}_1 = 0$, (2.5) yields

$$\begin{aligned} \tilde{L}_2\varphi(\varepsilon) + \sum_{n=3}^{\infty} \varepsilon^{n-2}\tilde{L}_n\varphi(\varepsilon) &= \frac{L(\varepsilon) - \lambda_0}{\varepsilon^2}P(\varepsilon)\varphi(\varepsilon) \\ &= \frac{\lambda(\varepsilon) - \lambda_0}{\varepsilon^2}\varphi(\varepsilon). \end{aligned}$$

Now taking the limit as $\varepsilon \to 0$ yields (2.16).

3. Parametric resonance. Let H be a complex Hilbert space, with scalar product (u,v) and norm $\|u\|$. Let A be the generator of a C_0 contraction semigroup $U(t)$ on H, $\|U(t)\| \leq 1$. Let Q be a bounded operator on H and let $p(t)$ be a continuous, real valued function. We assume

$$u \in D(A) \iff \bar{u} \in D(A), \qquad \overline{Au} = A\bar{u}, \qquad \overline{Qu} = Q\bar{u}, \tag{3.1}$$

and

$$p \text{ has period } T \text{ with } \int_0^T p(t)dt = 0. \tag{3.2}$$

We consider the abstract differential equation for a function $t \to u(t)$ taking values in H,

$$\frac{du}{dt} = Au + \varepsilon p(t)Qu, \tag{3.3}$$

where ε is a small parameter.

We assume that $\mu = -\kappa_0 + i\nu_0$, with $\kappa_0 \geq 0$ and $\nu_0 > 0$, is an eigenvalue of A with geometric and algebraic multiplicity $m \geq 1$. Thus there are independent eigenvectors

$$A\varphi_j = \mu\varphi_j, \qquad j = 1, \dots, m. \tag{3.4}$$

Because of (3.1), $\bar{\mu}$ is also an eigenvalue for A, with eigenvectors $\bar{\varphi}_j$:

$$A\bar{\varphi}_j = \bar{\mu}\bar{\varphi}_j, \qquad j = 1, \dots, m.$$

Finally we assume

$$\nu_0 = 2\pi/T \tag{3.5}$$

and we assume that

$$\lambda_0 = e^{\mu T} = e^{-\kappa_0 T} \tag{3.6}$$

is an isolated point of the spectrum of $U(T)$.

It follows that λ_0 is an eigenvalue of $U(T)$ of algebraic and geometric multiplicity $2m$. Let E denote the $2m$ dimensional eigenspace.

The solution of the initial value problem for (3.3) is given by an evolution operator $U_\varepsilon(t,s)$ for $s \leq t$; see Kato [5]. Integrating (3.3), we have

$$U_\varepsilon(t,s) = U(t-s) + \varepsilon \int_s^t U(t-\tau)p(\tau)QU_\varepsilon(\tau,s)d\tau \quad \text{for} \quad s \leq t. \tag{3.7}$$

THEOREM 3.1. *$\varepsilon \to U_\varepsilon(t,s)$ is an entire holomorphic family of bounded evolution operators $U_\varepsilon(t,s) : H \to H$ for $s \leq t$.*

Proof. We compute the derivatives of $U_\varepsilon(t,s)$ at $\varepsilon = 0$ formally, and show that these derivatives yield a convergent power series. For $f \in H$ and s fixed, let

$$z(t,\varepsilon) = U_\varepsilon(t,s)f, \qquad t \geq s.$$

Then $z(t,\varepsilon)$ satisfies

$$z(t,\varepsilon) = U(t-s)f + \varepsilon \int_s^t U(t-\tau)p(\tau)Qz(\tau,\varepsilon)d\tau.$$

Formally differentiating, we see that

$$\frac{\partial z(t,\varepsilon)}{\partial\varepsilon} = \int_s^t U(t-\tau)p(\tau)Qz(\tau,\varepsilon)d\tau + \varepsilon\int_s^t U(t-\tau)p(\tau)Q\frac{\partial z(\tau,\varepsilon)}{\partial\varepsilon}d\tau, \tag{3.8}$$

so that

$$z_1(t) \equiv \frac{\partial z(t,0)}{\partial\varepsilon} = \int_s^t U(t-\tau)p(\tau)Qz_0(\tau)d\tau, \tag{3.9}$$

where $z_0(\tau) = U(\tau)f$. In general we have

$$z_n(t) \equiv \frac{\partial^n z(t,0)}{\partial\varepsilon^n} = n\int_s^t U(t-\tau)p(\tau)Qz_{n-1}(\tau)d\tau. \tag{3.10}$$

Let $\alpha = \max|p(t)|$. Equation (3.10) yields the inequality

$$\begin{aligned}\|z_n(t)\| &\le n\alpha\int_s^t \|U(t-\tau)\|\|Q\|\|z_{n-1}(\tau)\|d\tau\\ &\le n\alpha\|Q\|\int_s^t \|z_{n-1}(\tau)\|d\tau.\end{aligned}$$

By induction we find that

$$\|z_n(t)\| \le (\alpha\|Q\|(t-s))^n\|f\|.$$

Thus $z(t,\varepsilon)$ has the convergent power series

$$z(t,\varepsilon) = \sum_0^\infty \frac{z_n(t)\varepsilon^n}{n!} \tag{3.11}$$

with

$$\|z(t,\varepsilon)\| \le \sum_0^\infty \frac{(\alpha\|Q\|(t-s)|\varepsilon|)^n}{n!}\|f\| \le e^{\alpha\|Q\|(t-s)|\varepsilon|}\|f\|.$$

The theorem is proved.

Now we wish to apply the results of section 2 to the holomorphic family of operators

$$\varepsilon \to L(\varepsilon) \equiv U_\varepsilon(T,0)$$

and investigate the behavior of the eigenvalues that split from λ_0 for $\varepsilon \ne 0$. From (3.9) and (3.10) we see that

$$L_1 f = \int_0^T U(T-s)p(s)QU(s)fds, \tag{3.12}$$

and

$$L_2 f = \int_0^T U(T-s)p(s)QU(s)\int_0^s U(s-\tau)p(\tau)QU(\tau)fd\tau ds. \tag{3.13}$$

To get more information about how the eigenvalue λ_0 splits, we use the $2m$ dimensional basis of eigenvectors of E, arranged as follows:

$$\varphi_1, \bar{\varphi}_1, \ldots, \varphi_m, \bar{\varphi_m}. \tag{3.14}$$

We introduce a basis for the $2m$ dimensional eigenspace E^* of $U(T)^*$ (with the same real eigenvalue $\lambda_0 = \exp(-\kappa_0 T)$). Let $\psi_1, \ldots, \psi_m$ be the eigenvectors of A^* with eigenvalue $\bar{\mu}$:

$$A^*\psi_j = \bar{\mu}\psi_j, \qquad j = 1, \ldots, m.$$

Then, because of (3.1), $\bar{\psi}_j$ are also eigenvectors of A^*

$$A^*\bar{\psi}_j = \mu\bar{\psi}_j, \qquad j = 1, \ldots, m.$$

We take the basis of E^* as $\psi_1, \bar{\psi}_1, \ldots, \psi_m, \bar{\psi}_m$. We observe that

$$(\varphi_j, \bar{\psi}_k) = 0, \qquad 1 \leq j, \qquad k \leq m, \tag{3.15}$$

whence

$$(\bar{\varphi}_j, \psi_k) = \overline{(\varphi_j, \bar{\psi}_k)} = 0, \qquad 1 \leq j, \qquad k \leq m.$$

This is easily seen since

$$\mu(\varphi_j, \bar{\psi}_k) = (A\varphi_j, \bar{\psi}_k) = (\varphi_j, A^*\bar{\psi}_k) = (\varphi_j, \mu\bar{\psi}_k) = \bar{\mu}(\varphi_j, \bar{\psi}_k).$$

Since we assume $\nu_0 = Im(\mu) > 0$, this implies (3.15).

Let N be the $2m \times 2m$ matrix consisting of the 2×2 blocks

$$N_{i,j} = \begin{bmatrix} (\varphi_i, \psi_j) & (\bar{\varphi}_i, \psi_j) \\ (\varphi_i, \bar{\psi}_j) & (\bar{\varphi}_i, \bar{\psi}_j) \end{bmatrix}, \qquad i, j = 1, \ldots, m. \tag{3.16}$$

We let M be the $2m \times 2m$ matrix consisting of the 2×2 blocks

$$M_{i,j} = \begin{bmatrix} (L_1\varphi_i, \psi_j) & (L_1\bar{\varphi}, \psi_j) \\ (L_1\varphi_i, \bar{\psi}_j) & (L_1\bar{\varphi}_i, \bar{\psi}_j) \end{bmatrix}, \qquad i, j = 1, \ldots, m. \tag{3.17}$$

Here L_1 is given by (3.12).

THEOREM 3.2. *Assume* (3.1) *and* (3.2). *Then the values $\rho_j = \lambda_j'(0)$ of the derivatives of the branches $\lambda_j(\varepsilon)$ are the roots of the characteristic equation*

$$\det(\rho N - M) = 0. \tag{3.18}$$

Furthermore, the roots ρ of (3.18) are symmetric with respect to the origin and with respect to the real axis.

Proof. From Theorem 2.3, we know that the derivatives $\lambda_j'(0)$ are the eigenvalues of $\tilde{L_1} = L_1PL_1$. Let w be an eigenvector of $\tilde{L_1}$ with eigenvalue ρ,

$$\rho w = \tilde{L_1}w = PL_1w, \tag{3.19}$$

because $w \in E$. Now take scalar products with ψ_j and $\bar{\psi}_j$ in (3.19),

$$\rho(w, \psi_j) = (PL_1w, \psi_j) = (L_1w, P^*\psi_j) = (L_1w, \psi_j) \tag{3.20}$$

and

$$\rho(w, \bar{\psi}_j) = (L_1 w, \bar{\psi}_j) \tag{3.21}$$

because $\psi_j, \bar{\psi}_j \in E^*$ and $P^* = I$ on E^*. Since $w \in E$, we can express w uniquely as

$$w = \sum_{j=1}^{m} a_{2j-1}\varphi_j + \sum_{j=1}^{m} a_{2j}\bar{\varphi}_j.$$

Substituting this expression into (3.20) and (3.21) yields

$$\rho N a = M a,$$

where $a = (a_1, a_2, \ldots, a_{2m})$. This is equivalent to (3.18).

By (3.15), the blocks

$$N_{i,j} = \begin{bmatrix} n_{i,j} & 0 \\ 0 & \bar{n}_{i,j} \end{bmatrix},$$

where $n_{i,j} = (\varphi_i, \psi_j)$.

To compute the elements of $M_{i,j}$, we use (3.12).

$$\begin{aligned}(L_1\varphi_i, \psi_j) &= \int_0^T (U(T-s)p(s)QU(s)\varphi_i, \psi_j)ds = \int_0^T p(s)(QU(s)\varphi_i, U(T-s)^*\psi_j)ds \\ &= \int_0^T p(s)e^{\mu s}(Q\varphi_i, e^{\bar{\mu}(T-s)}\psi_j) = e^{\mu T}(Q\varphi_i, \psi_j)\int_0^T p(s)ds = 0\end{aligned}$$

by (3.2). Because $U(t)$ and Q take real vectors into real vectors,

$$(L_1\bar{\varphi}_i, \bar{\psi}_j) = \overline{(L_1\varphi_i, \psi_j)} = 0.$$

Furthermore,

$$\begin{aligned} m_{i,j} \equiv (L_1\bar{\varphi}_i, \psi_j) &= e^{\mu T}(Q\bar{\varphi}_i, \psi_j)\int_0^T p(s)e^{(\bar{\mu}-\mu)s}ds \\ &= e^{-\kappa_0 T}(Q\bar{\varphi}_i, \psi_j)p_2, \end{aligned} \tag{3.22}$$

where

$$p_2 = \int_0^T p(t)e^{-2i\nu_0 t}dt.$$

Also $(L_1\varphi_i, \bar{\psi}_j) = \overline{(L_1\bar{\varphi}_i, \psi_j)}$. Thus the 2×2 blocks $M_{i,j}$ have the form

$$M_{i,j} = \begin{bmatrix} 0 & m_{i,j} \\ \bar{m}_{i,j} & 0 \end{bmatrix}.$$

The matrix $\rho N - M$ consists of the 2×2 blocks

$$\rho N_{i,j} - M_{i,j} = \begin{bmatrix} \rho n_{i,j} & -m_{i,j} \\ -\bar{m}_{i,j} & \rho\bar{n}_{i,j} \end{bmatrix}.$$

Let $l(\rho) = \det(\rho N - M)$, whence $\overline{l(\rho)} = \det(\bar\rho \bar N - \bar M)$ with blocks

$$\bar\rho \bar N_{i,j} - \bar M_{i,j} = \begin{bmatrix} \bar\rho \bar n_{i,j} & -\bar m_{i,j} \\ -m_{i,j} & \bar\rho n_{i,j} \end{bmatrix}.$$

Now interchange rows $2i-1$ and $2i$, $i = 1, \ldots, m$ and columns $2j-1$ and $2j$, $j = 1, \ldots, m$ of $\bar\rho \bar N - \bar M$. These row and column interchanges yield $\bar\rho N - M$. Hence

$$\overline{l(\rho)} = \det(\bar\rho \bar N - \bar M) = \det(\bar\rho N - M) = l(\bar\rho).$$

Thus $l(\rho)$ has real coefficients so that the roots come in conjugate pairs.

Next we show that $l(-\rho) = l(\rho)$. In fact $l(-\rho) = \det(-\rho N - M)$ and $-\rho N - M$ has the 2×2 blocks

$$\begin{bmatrix} -\rho n_{i,j} & -m_{i,j} \\ -\bar m_{i,j} & -\rho \bar n_{i,j} \end{bmatrix}.$$

Multiply the odd numbered rows by -1 and the even numbered columns by -1. These row and column operations on $-\rho N - M$ yield $\rho N - M$. Hence

$$l(-\rho) = \det(-\rho N - M) = \det(\rho N - M) = l(\rho).$$

This completes the proof of Theorem 3.2.

Remark. Let $\lambda_j(\varepsilon)$ be a branch of the eigenvalues splitting from λ_0, with eigenvector $\varphi_j(\varepsilon)$. Then the first term in an asymptotic expansion of $\varphi_j(\varepsilon)$ is an eigenvector of $\lambda_j'(0)N - M$.

When $\tilde L_1 = 0$, we see in (2.13) that

$$\tilde L_2 = PL_2P - PL_1SL_1P,$$

where S is the reduced resolvant: Let K be the $2m \times 2m$ matrix consisting of the 2×2 blocks

$$(3.23)\quad K_{i,j} = \begin{bmatrix} (L_2\varphi_i - L_1SL_1\varphi_i, \psi_j) & (L_2\bar\varphi_i - L_1SL_1\bar\varphi_i, \psi_j) \\ (L_2\varphi_i - L_1SL_1\varphi_i, \bar\psi_j) & (L_2\bar\varphi_i, -L_1SL_1\bar\varphi_i, \bar\psi_j) \end{bmatrix}, \quad i, j = 1, \ldots, m.$$

THEOREM 3.3. *Assume* (3.1) *and* (3.2). *Suppose that* $p_2 = 0$ *and that* $\det(N) \neq 0$. *Then* $\tilde L_1 = 0$ *so that* $\lambda_j'(0) = 0$ *for each of the branches* $\lambda_j(\varepsilon)$. *In this case,* $\lambda_j(\varepsilon)$ *is twice differentiable at* $\varepsilon = 0$ *and* $\rho_j = 2\lambda_j''(0)$ *are the roots of the characteristic equation*

$$(3.24)\qquad \det(\rho N - K) = 0.$$

The roots of (3.24) are symmetric with respect to the real axis.

Proof. Since N is assumed to be invertible, the matrix for $\tilde L_1 = PL_1P$ in the basis (3.14) for the eigenspace E is $N^{-1}M$. In fact, for $f \in H$, the projection can be expressed

$$Pf = \sum_{j=1}^{m} c_{2j-1}\varphi_j + \sum_{j=1}^{m} c_{2j}\bar\varphi_j,$$

where

$$c = N^{-1}((f, \psi_1), (f, \bar\psi_1), \ldots, (f, \psi_m), (f, \bar\psi_m)).$$

Now taking $f = L_1 g$ with

$$g = \sum_{j=1}^{m} a_{2j-1}\varphi_j + \sum_{j=1}^{m} a_{2j}\bar{\varphi}_j$$

we see that

$$((f,\psi_1),(f,\bar{\psi}_1),\ldots,(f,\psi_m),(f,\bar{\psi}_m)) = Ma.$$

Thus the coordinates of $\tilde{L}_1 g = PL_1 g$ are related to the coordinates of g by $c = N^{-1}Ma$.

Now assuming $p_2 = 0$, we see by (3.22) that $M = 0$, whence $\tilde{L}_1 = 0$. Thus $\lambda_j'(0) = 0$ for all branches of $\lambda_j(\varepsilon)$. By Theorem 3.1, λ_j is twice differentiable at $\varepsilon = 0$, and for each branch, $\rho_j = 2\lambda_j''(0)$ is an eigenvalue of $\tilde{L}_2$. If w is an eigenvector of $\tilde{L}_2$ with eigenvalue ρ,

$$\rho w = \tilde{L}_2 w,$$

we take scalar product with ψ_j and $\bar{\psi}_j$ to obtain

$$\rho(w,\psi_j) = (\tilde{L}_2,\psi_j)$$

$$= (PL_2 w - PL_1 S L_1 w, \psi_j) = (L_2 w - L_1 S L_1 w, \psi_j)$$

and similarly for $\bar{\psi}_j$. Then writing $w = \sum a_{2j-2}\varphi_j + a_{2j}\bar{\varphi}_j$, we deduce, as in the proof of Theorem 3.2, that $\rho N a = K a$ where K is given by (3.23). Finally we note that because L and S take real vectors into real vectors, the blocks $K_{i,j}$ have the form

$$K_{i,j} = \begin{bmatrix} k_{i,j} & l_{i,j} \\ \bar{l}_{i,j} & \bar{k}_{i,j} \end{bmatrix}.$$

It follows easily that if ρ satisfies (3.24), then so does $\bar{\rho}$.

Remark. The proof of Theorems 3.2 and 3.3 actually depends only on (3.7). Once this formula is established, we do not need to know the differential equation solved by $U_\varepsilon(t,s)$.

4. Scattering frequencies. We apply the results of sections 2 and 3 to the wave equation in three space dimensions with a time-periodic potential. Let $q_0(x) \in L^\infty(R^3)$, $x \in R^3$, with $q_0(x) \geq 0$ and $q_0(x) = 0$ for $|x| > 1$. We assume that $q_0(x) > 0$ on some open subset. Next let $q_1(x) \in L^\infty(R^3)$ be real valued with $q_1(x) = 0$ for $|x| > 1$. We consider the wave equations

$$u_{tt} - \Delta u + q_0(x)u = 0 \tag{4.1}$$

and

$$u_{tt} - \Delta u + q_0(x)u + \varepsilon p(t) q_1(x) u = 0, \tag{4.2}$$

where $p(t)$ is real valued, continuous, and has period T with

$$\int_0^T p(t)dt = 0. \tag{4.3}$$

We write (4.1) and (4.2) as systems

$$u_t = Au \tag{4.4}$$

and

$$u_t = Au + \varepsilon p(t) Q u, \tag{4.5}$$

where now u is a pair, $u = [u(x,t), u_t(x,t)]$. A is the matrix differential operator

$$A = \begin{bmatrix} 0 & 1 \\ \Delta - q_0 & 0 \end{bmatrix} \tag{4.6}$$

and

$$Q = \begin{bmatrix} 0 & 1 \\ q_1 & 0 \end{bmatrix}. \tag{4.7}$$

The finite energy space H for (4.4) and (4.5) is the space of pairs $f = [f_1, f_2]$ which is the closure of $C_0^\infty(R^3) \times L^2(R^3)$ in the energy norm

$$\|f\| = \left[\int_{R^3} [|\nabla f_1|^2 + q_0 |f_1|^2] dx + \int_{R^3} |f_2|^2 dx\right]^{1/2}.$$

The scalar product on H is

$$(f, g) = \int_{R^3} [\nabla f_1 \cdot \nabla \bar{g}_1 + q_0 f_1 \bar{g}_1 + f_2 \bar{g}_2] dx.$$

It is well known that A generates a unitary group $U(t) : H \to H$. Applying Theorem 3.1, we see that the finite energy solutions of (4.5) are given by an evolution operator $V_\varepsilon(t, s) : H \to H$, and that $\varepsilon \to V_\varepsilon(t, s)$ is an entire function with values in the space of bounded operators on H.

We shall not apply the results of section 3 directly to (4.4) and (4.5). In fact, A has no eigenvalues. Instead we study the behavior of the solutions in a neighborhood of the support of the potential. In [6] Lax and Phillips developed a framework to treat this situation. We recall some elements of that framework. First we make a decomposition of the space of data H into subspaces which represent solutions which have left the region of the potential (outgoing), those which have not reached the region of the potential (incoming), and the remainder which interacts with the potential. The *outgoing subspace* is

$$D^+ = \{f \in H : U_0(t) f = 0 \text{ in } |x| \le t + 1, t > 0\}$$

and the *incoming subspace* is

$$D^- = \{f \in H : U_0(t) f = 0 \text{ in } |x| \le |t| + 1, t < 0\}.$$

Here $U_0(t)$ is the unitary group of the free wave equation (with $q_0 = 0$). We can decompose H as an orthogonal direct sum

$$H = D^+ \oplus K \oplus D^-. \tag{4.8}$$

Let P_+ be the orthogonal projection on the *orthogonal complement* of D^+, which is $D^- \oplus K$. P_- will denote the orthogonal projection onto $D^+ \oplus K$. Note that for $f \in D^\pm$, $f = 0$ in $|x| < 1$. Hence for any $f \in H$,

$$P_\pm f = f \quad \text{for} \quad |x| < 1. \tag{4.9}$$

Next we introduce the localized semigroup and evolution operator. We define

$$Z(t) = P_+ U(t) P_- \qquad \text{for} \quad t \geq 0. \tag{4.10}$$

Because $U(t)D^+ \subset D^+$ for $t \geq 0$, $Z(t)f = 0$ for $f \in D^\pm$. For $f \in K$, $Z(t)f \in K$. $Z(t)$ is in fact a contraction semigroup on K. We denote its generator by B, $Z(t) = \exp(tB)$.

Singularities of the solutions of (4.1) propagate with speed one and are unimpeded by the potential. Hence for $f \in H$, $U(t)f$ is smooth in $\{|x| < 1\}$ for $|t| > 2$. This implies that the spectrum of B is discrete, consisting of eigenvalues μ_j, with $Re(\mu_j) < 0$. Furthermore, any vertical strip in the complex plane contains at most a finite number of eigenvalues. The *scattering frequencies* for (4.1) are the complex numbers

$$\sigma_j = \nu_j + i\kappa_j$$

such that $i\sigma_j = \mu_j$. Note that $\kappa_j > 0$.

Because $Z(t)$ is compact, the eigenspace associated with each eigenvalue μ_j is finite dimensional. Assuming there are no generalized eigenvectors, solutions of (4.1) with data of compact support have an asymptotic expansion

$$u(x,t) \approx \sum_j e^{i\sigma_j t} v_j(x)$$

which approximates u in the local energy norm. The v_j are outgoing scattering eigensolutions of

$$-\Delta v + q_0(x)v = \sigma^2 v.$$

They satisfy the outgoing Sommerfeld radiation condition. The pair $[v_j, \mu_j v_j]$ does not belong to H because v_j grows exponentially as $|x| \to \infty$. However, for $|x| < 1$, the pair $[v_j, \mu_j v_j]$ agrees with an eigenvector φ_j of B with eigenvalue μ_j and

$$Z(t)\varphi_j = e^{\mu_j t}\varphi_j. \tag{4.11}$$

Now we define the localized evolution operator. Let

$$Z_\varepsilon(t,s) = P_+ V_\varepsilon(t,s) P_-.$$

Lemma 4.1. *$Z_\varepsilon(t,s)$ is a holomorphic family of evolution operators on K. Z_ε and Z satisfy the integral relation* (3.7).

Proof. That $Z_\varepsilon(t,s)$ is an evolution operator on K was shown in [1]. The holomorphic property is immediate because $\varepsilon \to V_\varepsilon(t,s)$ is holomorphic.

We need only verify that Z_ε and Z satisfy the integral relation (3.7). However, $U(t)$ and $V_\varepsilon(t,s)$ do satisfy this relation:

$$V_\varepsilon(t,s) = U(t-s) + \varepsilon \int_0^t U(t-\tau)p(\tau)QV_\varepsilon(\tau,s)ds.$$

Apply the projection P_+ to each term in this equation. Using the definitions of Z and Z_ε we find that for $f \in K$,

$$Z_\varepsilon(t,s)f = Z(t-s)f + \varepsilon \int_s^t Z(t-\tau)p(\tau)QV_\varepsilon(\tau,s)f d\tau.$$

But $QV_\varepsilon(\tau,s)f = QP_+V_\varepsilon(\tau,s)f = QZ_\varepsilon(\tau,s)f$ because $P_+g = g$ for $|x| < 1$. The lemma is proved.

Finally, after this lengthy preparation, we can apply the results of sections 2 and 3 to the semigroup $Z(t)$ and the evolution operator $Z_\varepsilon(t,s)$.

Let $\sigma_0 = \nu_0 + i\kappa_0$ be a scattering frequency of (4.1) with $\nu_0 > 0$. Thus $\mu_0 = i\sigma_0$ is an eigenvalue of B. $\lambda_0 = \exp(\mu_0 T) = \exp(i\sigma_0 T) = \exp(-\kappa_0 T) > 0$ is an eigenvalue of $Z(T)$.

Theorem 4.2. *Assume that the eigenvalue μ_0 has geometric and algebraic multiplicity m. Choose $T = 2\pi/\nu_0$, and assume that p is real valued, continuous, and satisfies* (3.2). *Then the scattering frequency σ_0 splits into s branches $\sigma_j(\varepsilon)$, $1 \le s \le m$. $\sigma_j(\varepsilon)$ is differentiable at $\varepsilon = 0$ and holomorphic in a punctured complex neighborhood of $\varepsilon = 0$.*

Proof. The general theory of section 2 applies here with $L(\varepsilon) = Z_\varepsilon(T,0)$. We need to verify that $\lambda_0 = \exp(\mu_0 T)$ is an isolated point of the spectrum of $L = L(0) = Z(T)$. But this follows because, as noted before, each vertical strip in the complex plane contains only a finite number of eigenvalues μ of B. This means that in each annulus centered at zero, $Z(T)$ has only a finite number of eigenvalues.

If $\lambda_j(\varepsilon)$ is a branch of the eigenvalues splitting from λ_0, then $\lambda_j'(0)$ exists. Now

$$\sigma_j(\varepsilon) = \frac{\log(\lambda_j(\varepsilon))}{iT}$$

are the branches of the scattering frequencies splitting from σ_0. They are holomorphic in a punctured neighborhood of $\varepsilon = 0$, and differentiable at $\varepsilon = 0$ with

$$\sigma_j'(0) = \frac{\lambda_j'(0)}{iT\lambda_0} = \frac{e^{\kappa_0 T}\lambda_j'(0)}{iT}. \tag{4.12}$$

This ends the proof of Theorem 4.2.

Corollary 4.3. *Let $\sigma_j(\varepsilon)$ be the branches of the scattering frequency which split from resonant scattering frequency σ_0. The $\sigma_j'(0)$ have complex values α such that if α is such a direction, then so is $-\bar{\alpha}$ and so is $-\alpha$.*

This is an immediate consequence of Theorem 3.2, the remark at the end of section 3, and (4.12).

Example 1. We specialize the discussion to the case where q_0 is a real constant, $q_0 > 0$ on $\{|x| < 1\}$, and $q_1(x) \equiv 1$ on $\{|x| < 1\}$. If σ_0 is a scattering frequency of (4.1) that corresponds to a spherically symmetric scattering eigenfunction, then the perturbed solution of (4.2) will also be spherically symmetric. Thus we restrict our attention to spherically symmetric solutions of (4.1) and (4.2). We make the change of dependent variable $z(r,t) = ru(r,t)$. z now satisfies

$$z_{tt} - z_{rr} = \begin{cases} -q_0 z, & 0 < r < 1, \\ 0, & r > 1, \end{cases}$$

$$z(0,t) = 0,$$

$$z(r,0) = f_1(r), \qquad z_t(r,0) = f_2(r).$$

Now if $f_1 = f_2 = 0$ for $r > 1$, then $z = z(t-r)$ for $t \geq 0$ and $r > 1$. Thus, in this case, the values of z are completely determined by the solution of the initial boundary value problem

$$z_{tt} - z_{rr} + q_0 z = 0, \qquad 0 < r < 1, \tag{4.13}$$

$$z(0,t) = 0, \qquad z_t + z_r = 0 \quad \text{on} \quad r = 1, \tag{4.14}$$

$$z(r,0) = f_1(r), \qquad z_t(r,0) = f_2(r).$$

Because $P_\pm f = f$ for $|x| < 1$, and $Z(t)f = P_+U(t)P_-f$, we see that when $f = 0$ for $|x| > 1$, $Z(t)f = U(t)f$ on $\{|x| < 1\}$. Consequently, in the case of spherical symmetry, the solutions of the localized evolution equation

$$\frac{dv}{dt} = Bv$$

are exactly the solutions of (4.13), (4.14), with $v = [z, z_t]$. The eigenfunctions of B are the spatial factors of the solutions of (4.13), (4.14) of the form $z(r,t) = \exp(\mu t)w(r)$ where w satisfies

$$-w_{rr} + q_0 + \mu^2 = 0,$$

$$w(0) = 0, \qquad \mu w(1) + w_r(1) = 0.$$

It is convenient to seek the solutions of this problem in the form $w(r) = \sin(\gamma r)$. γ and μ must satisfy

$$\gamma^2 + \mu^2 + q_0 = 0, \qquad \mu \sin(\gamma) + \gamma \cos(\gamma) = 0. \tag{4.15}$$

Writing μ in terms of a scattering frequency, $\mu = i\sigma_0$, we see that for large q_0, γ and σ_0 have the asymptotic approximations

$$\gamma(q_0) = n\pi(1 + i/\sqrt(q_0) - 1/q_0) + O(q_0^{-3/2}),$$

$$\sigma_0(q_0) = \sqrt{\gamma^2 + q_0}.$$

We have taken the roots γ and σ_0 in the first quadrant. There is another root $\bar{\gamma}$ and corresponding root $-\bar{\sigma_0}$, which corresponds to $\bar{\mu}$.

In this case, the eigenvalue $\mu = i\sigma_0$ is simple ($m = 1$). The eigenfunctions of B and B^* are

$$\varphi(r) = \begin{bmatrix} \sin(\gamma r)/r \\ i\sigma_0 \sin(\gamma r)/r \end{bmatrix} \quad \text{and} \quad \psi(r) = \begin{bmatrix} \sin(\bar{\gamma})/r \\ i\bar{\sigma}_0 \sin(\bar{\gamma} r)/r \end{bmatrix}, \qquad 0 < r < 1.$$

The matrices M and N are 2×2 and the characteristic equation $\det(\rho N - M) = 0$ has the roots

$$\rho = \pm\frac{|m_{1,1}|}{|n_{1,1}|}.$$

Here

$$n_{1,1} = (\varphi, \psi), \qquad \text{and} \quad m_{1,1} = e^{-\kappa_0 T} p_2(Q\bar{\varphi}, \psi).$$

Thus the scattering frequency σ_0 splits into two branches with

$$\sigma'(0) = \pm i\frac{|p_2(Q(\bar{\varphi}, \psi)|}{T|(\varphi, \psi)|}.$$

Remark. Computations show that when $p_2 \neq 0$, the branches of the resonant scattering frequency move vertically on the line $Re(\sigma) = Re(\sigma_0)$ as $\varepsilon \uparrow$, and eventually one of the branches crosses the real axis. This means that for $\varepsilon > 0$ sufficiently large, there are outgoing eigensolutions of the perturbed problem (4.2) that grow exponentially as $t \to \infty$. We hope to prove this in the future.

REFERENCES

[1] J. Cooper and W. Strauss, *Abstract scattering theory for time-periodic systems with applications to electromagnetism*, Indiana Univ. Math. J., 34 (1985), pp. 33–83.

[2] J. Cooper, G. Perla-Menzala, and W. Strauss, *On the scattering frequencies of time-dependent potentials*, Math. Methods Appl. Sci., 8 (1986), pp. 576–584.

[3] J. Howland, *Quantum stability in Schroedinger operators: The quantum mechanical many-body problem,* in Proceedings of a Workshop held at Aarhus, Denmark, Lecture Notes in Phys. 403, Springer-Verlag, Berlin, 1992, pp. 100–122.

[4] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[5] T. Kato, *Linear evolution equations of "hyperbolic" type*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 17 (1970), pp. 241–258.

[6] P. Lax and R. Phillips, *Scattering Theory*, Academic Press, New York, 1967.

[7] G. Majda and M. Wei, *Numerical computation of the scattering frequencies for a cylindrically symmetric potential*, SIAM J. Sci. Comput., 14 (1993), pp. 295–309.

[8] G. Majda, M. Wei, and W. Strauss, *Numerical computation of the scattering frequencies for acoustic wave equations*, J. Comput. Phys., 75 (1988), pp. 345–358.

[9] J. J. Stoker, *Nonlinear Vibrations in Mechanical and Electrical Systems*, Interscience, New York, 1950.

SIAM J. Math. Anal.
Vol. 31, No. 4, pp. 836–861

LOW-FREQUENCY ELECTROMAGNETIC SCATTERING*

H. AMMARI† AND J.-C. NÉDÉLEC†

Abstract. The main result of this paper is to reduce the calculation of higher-order terms in the asymptotic expansions of the electric and magnetic fields at low frequencies to the solutions of certain canonical problems. Our approach is based on coupling the power series representation of the scattered fields with expansion of the exact nonlocal radiation condition. We also provide a new and simple variational proof of the convergence of the electric and magnetic fields solutions of the scattering problem for the Maxwell equations as the frequency goes to zero. Besides its theoretical interest, our analysis is motivated by its application to the numerical computation of the higher-order terms. These higher-order terms may be combined to Padé approximations to enlarge the domain of applicability of the low-frequency scattering to predict more accurately the reponse of diffraction problems for heteregeneous Maxwell's equations in the resonance region where the wavelength and the dimension of the dielectric material are of the same order.

Key words. low-frequency, Maxwell's equations

AMS subject classifications. 35C20, 35B40, 35Q60

PII. S0036141098343604

1. Introduction and statement of the problem.

The scattering of electromagnetic waves from bounded objects whose dimensions are small compared with the length of the incident wave has been the subject of considerable study for more than a century. This problem is of interest in geophysics, astrophysics, electrical engineering, physics of the atmosphere and ocean, medicine, biology, and other fields. The study of wave scattering at low frequencies was pioneered by Rayleigh [33], who continued this work until his death. His contributions in this area provide the foundation on which almost all subsequent work is based. A low-frequency asymptotic for Maxwell's boundary-value problem and transmission problem has been given, for instance, by Müller and Niemeyer [26], Stevenson [34], Kleinman [19], Werner [41] and [39], Picard [29], Kleinman–Senior [20], Weck and Witsch [38], Kress [21], Ramm et al. [32], Ramm [30], and Ramm–Somersalo [31]. Kleinman and Senior systematized the calculation of the dominant term in the low-frequency limit for electromagnetic problems involving impenetrable, penetrable, nonlossy, and lossy obstacles. With boundary integral methods, Werner [40], [42], [41], [39] and Kress [21] obtained the limit of the solution to the scattering problem by a perfectly conducting object for Maxwell's equations. Kriegsmann and Reiss [22] used matched asymptotic expansions to align together the local and far-field approximations at low frequencies.

One of our main motivations in this paper is to reduce the calculation of the higher-order terms which are the Taylor coefficients in the asymptotic expansion of the electric and magnetic fields with respect to the frequency. By combining these higher-order terms and Padé approximations, we may enlarge the domain of applicability of the low-frequency scattering to predict more accurately the response of diffraction problems for heterogeneous Maxwell's equations in the resonance region where the wavelength and the dimension of the dielectric material are of the same order. We

*Received by the editors August 17, 1998; accepted for publication (in revised form) May 3, 1999; published electronically April 4, 2000.

http://www.siam.org/journals/sima/31-4/34360.html

†Centre de Mathématiques Appliquées, CNRS UMR 7641, École Polytechnique, 91128 Palaiseau Cedex, France (ammari@cmapx.polytechnique.fr, nedelec@cmapx.polytechnique.fr).

may also gain insight into the spectrum of singularities of the fields as functions of the frequency: the set of resonances. This program is in the spirit of "the method of variation of boundaries" introduced recently by Bruno and Reitich [6], [7], [8], [9], [10], [11]. Their new method is very attractive. Many numerical experiments show that it has several orders of magnitude more accurate than other classical methods such as integral or variational formalisms. The implementation of our "null-frequency method" is in progress and numerical results in the resonance region for Maxwell's equations as well as the Helmholtz equation will be published in a forthcoming paper.

In the present paper we restrict ourselves to presenting, in some detail, the derivation and well-posedness of higher-order terms and the proof of convergence of the electric and magnetic fields as the frequency goes to zero by a simple variational method.

In fact, our present work is closely related to the work of Picard [29] who proved the convergence for nonhomogeneous Maxwell's equations using Hilbert space concepts. The approach of Picard [29] is based on the limiting absorption principle for a modified first-order Maxwell system to overcome the fact that the case $\omega = 0$ is embedded in the continuous spectrum of the modified first-order Maxwell operator. He reformulated Maxwell's equations in the following way:

$$\Big(M + N - \omega\Big)\, U^\omega \;=\; F^\omega,$$

where

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \dfrac{i}{\varepsilon}\,\overrightarrow{\mathrm{curl}} & 0 \\ 0 & \dfrac{i}{\mu}\,\overrightarrow{\mathrm{curl}} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & i\,\mathrm{div}\,\varepsilon & 0 & 0 \\ i\,\overrightarrow{\mathrm{grad}} & 0 & 0 & 0 \\ 0 & 0 & 0 & i\,\overrightarrow{\mathrm{grad}} \\ 0 & 0 & i\,\mathrm{div}\,\mu & 0 \end{pmatrix},$$

and

$$U^\omega = \begin{pmatrix} 0 \\ E^\omega \\ H^\omega \\ 0 \end{pmatrix}.$$

Here ε is the electric permittivity, μ the magnetic permeability, E^ω the electric field, H^ω the magnetic field, and F^ω a given vector function with a bounded support. His method cannot be easily extended to obtain the rate of the convergence of the electric and magnetic fields with higher-order terms since the assumption of the support of F^ω is crucial for his proof (Theorem 3, p. 70). In none of the works cited above is a variational method for calculating the higher-order terms given or convergence with these terms analyzed. This is our basic aim in the present paper. We also provide a new and simple variational proof of the convergence of the electric and magnetic fields as the frequency goes to zero. Our approach is based on coupling the power series representation of the scattered fields with expansion of the exact nonlocal radiation condition. To the best of our knowledge, this approach is new; it is different from those mentioned above.

We first formulate the scattering problem equivalently on a ball B_R of radius R containing the inhomogeneity by making use of an adequate Steklov–Poincaré operator on the sphere $S_R = \partial B_R$ called, throughout what follows, the electromagnetic operator. Either the expression of this pseudodifferential operator or the variational formulation used here for Maxwell's equations is now well known (see, for instance, [17] and [27]). Then, the electric and magnetic fields, as well as the electromagnetic operator, are expanded in a power series with respect to the frequency. Making use of some properties of the Hankel functions, we show that in the asymptotic expansion of the electromagnetic operator with respect to the frequency, all the coefficients which are pseudodifferential operators are of an order less than the order of the electromagnetic operator. This important result is an essential tool which will permit us to derive the appropriate boundary conditions satisfied by the higher order terms of the electric and magnetic fields on the sphere S_R. These boundary conditions on the fictive surface S_R complete the reduction of the calculation of the higher-order terms in the asymptotic expansions of the electric and magnetic fields to two canonical problems which are uniquely solvable. A variational proof of the convergence with these higher-order terms will be given in the last part of this paper. Let us note that in addition to the difficulty coming from the asymptotic expansion of the electromagnetic operator, there is another difficulty that is due to a lack of coerciveness in the variational formulation. This difficulty is overcome by using a compactness result (see, for example, Lemma 5.2).

The extension of this approach to handle scattering problems by dielectric objects containing conducting bodies is presented in [4] where the dependence of the asymptotics on the topological properties of the conductors is completely shown.

Let us now introduce the scattering problem. Suppose that a bounded inhomogeneity characterized by permittivity ε and permeability μ is illuminated by a time-harmonic electromagnetic wave given by $(E^{in}_\omega, H^{in}_\omega)$ (where E^{in}_ω is the electric field and H^{in}_ω is the magnetic field). The incoming wave $(E^{in}_\omega, H^{in}_\omega)$ is assumed to be a classical solution of the Maxwell system

$$\begin{cases} \overrightarrow{\text{curl}}\, H^{in}_\omega \;=\; -i\omega\varepsilon_0\, E^{in}_\omega, \\ \overrightarrow{\text{curl}}\, E^{in}_\omega \;=\; i\omega\mu_0\, H^{in}_\omega \end{cases}$$

in all of R^3 except possibly for a finite number of points located outside a sphere strictly containing the inhomogeneity. This incoming field will interact with the inhomogeneity to give rise to a scattered field $(E^{sc}_\omega, H^{sc}_\omega)$. Let ε, μ be two functions of the spatial variable x; ε_0 and μ_0 two positive constants; and $r = |x|$. The total field (E^ω, H^ω) satisfies the time-harmonic Maxwell system

$$\begin{cases} \overrightarrow{\text{curl}}\, H^\omega \;=\; -i\omega\varepsilon\, E^\omega \quad \text{in } R^3, \\ \overrightarrow{\text{curl}}\, E^\omega \;=\; i\omega\mu\, H^\omega \quad \text{in } R^3, \\ H^\omega \;=\; H^{sc}_\omega + H^{in}_\omega \quad \text{in } R^3, \\ E^\omega \;=\; E^{sc}_\omega + E^{in}_\omega \quad \text{in } R^3, \\ \lim_{r\to+\infty}\left(\sqrt{\mu_0}\, H^{sc}_\omega \wedge x - r\sqrt{\varepsilon_0}\, E^{sc}_\omega\right) \;=\; 0. \end{cases} \tag{1.1}$$

We assume that the electric permittivity and the magnetic permeability, ε and μ, respectively, are in $L^\infty(R^3)$. In the next section, we will add some restrictive conditions regarding the regularity of these coefficients to ensure the uniqueness of a solution to (1.1). We also assume that the inhomogeneity is bounded so that there exists a constant $R_0 > 0$ such that $\varepsilon(x) = \varepsilon_0$ and $\mu(x) = \mu_0$ if $r = |x| \geq R_0$. The support of the function $\varepsilon - \varepsilon_0$, denoted by $\mathrm{Supp}(\varepsilon - \varepsilon_0)$, and that of $\mu - \mu_0$ are included in the ball $B_{R_0} = \{|x| < R_0\}$.

For some fixed $R > R_0$ we define the ball $B_R = \{x \in R^3, |x| < R\}$ and the sphere $S_R = \{x \in R^3, |x| = R\}$.

The present study is concerned with the analysis of the behavior of the fields (E^ω, H^ω) solutions of (1.1) as the frequency ω tends to zero. Our aim in this paper is to prove the convergence of the fields (E^ω, H^ω) and to reduce the calculation of higher-order terms to solutions of certain canonical problems. A similar program was carried out by the authors (with Laouadi) for the scattering problem from a small conductor embedded in a homogeneous chiral media [3]. For this problem, the authors employed representations of electric and magnetic fields to reformulate the scattering problem as an integral equation over the scattering surface and to represent the fields in terms of surface fields. Using a Hodge decomposition of the tangent fields, the authors characterized the dependence of the asymptotics on the topological properties of the scatterer and have shown the effect of the chirality admittance. The extension of the present study to the scattering of electromagnetic waves in a chiral media from a small inhomogeneity does not lead to any specific difficulty apart from much more cumbersome notations and calculations.

Now we introduce some standard notation. For any smooth vector field w, we denote by w_{S_R} its tangential component on S_R:

$$w_{S_R} = -n \wedge (n \wedge w) \quad \text{on } S_R.$$

For any smooth function f or any vector field u defined on S_R, we get an extension of each one to $R^3 \setminus \overline{B_R}$ by setting

$$\forall\, x \in R^3 \setminus \overline{B_R}, \quad \tilde{f}(x) = f\left(\frac{x}{|x|}\right), \quad \tilde{u}(x) = u\left(\frac{x}{|x|}\right).$$

We also introduce the following boundary differential operators:

the tangential gradient of a function f: $\overrightarrow{\mathrm{grad}}_{S_R} f = \overrightarrow{\mathrm{grad}}\, \tilde{f}|_{S_R}$,

the surface divergence of a field u: $\mathrm{div}_{S_R} u = \mathrm{div}\, \tilde{u}|_{S_R}$,

the vector rotational of a function: $\overrightarrow{\mathrm{curl}}_{S_R} f = \overrightarrow{\mathrm{curl}}\, (fn)|_{S_R}$,

the scalar rotational of a vector: $\mathrm{curl}_{S_R} u = n \,.\, \overrightarrow{\mathrm{curl}}\, \tilde{u}|_{S_R}$,

the Laplace–Beltrami operator on S_R defined on scalar functions by

$$\Delta_{S_R} f = \mathrm{div}_{S_R} \overrightarrow{\mathrm{grad}}_{S_R} f = -\mathrm{curl}_{S_R} \overrightarrow{\mathrm{curl}}_{S_R} f.$$

To state our boundary-value problem in a suitable mathematical form, we shall use the following notation for the usual functional spaces [15], [35]:

$\mathcal{D}'(B_R)$ is the space of distributions in B_R.

$L^2(B_R)$ is the space of complex square integrable functions defined in B_R.

$L^\infty(R^3)$ is the space of bounded functions defined in R^3.

$H^1(B_R) = \{\varphi \in L^2(B_R), \overrightarrow{\mathrm{grad}}\, \varphi \in (L^2(B_R))^3\}$.

$H^1_0(B_R) = \{\varphi \in H^1(B_R), \varphi = 0 \,\mathrm{on}\, S_R\}$.

$H(\mathrm{curl}, B_R) = \{u \in (L^2(B_R))^3, \overrightarrow{\mathrm{curl}}\, u \in (L^2(B_R))^3\}$.

$H_0(\mathrm{curl}, B_R) = \{\ u \in H(\mathrm{curl}, B_R), u \wedge n = 0 \text{ on } B_R\ \}$.

$TL^2(S_R) = \{c \in (L^2(S_R))^3, c \,.\, n = 0\}$.

$TH^s(S_R) = \{c \in (H^s(S_R))^3, c\,.\,n = 0\}$.

$TH^{-1/2}(\mathrm{curl}, S_R) = \{c \in (H^{-1/2}(S_R))^3, c\,.\,n = 0, \mathrm{curl}_{\partial\Omega}\, c \in H^{-1/2}(S_R)\}$.

$TH^{-1/2}(\mathrm{div}, S_R) = \{c \in (H^{-1/2}(S_R))^3, c\,.\,n = 0, \mathrm{div}_{S_R}\, c \in H^{-1/2}(S_R)\}$.

$TH^{-1/2}(\mathrm{curl}, S_R)$ and $TH^{-1/2}(\mathrm{div}, S_R)$ are Hilbert spaces. Finally, we recall without proof the following well-known duality result due to Paquet [28]: $TH^{-1/2}(\mathrm{curl}, S_R) = (TH^{-1/2}(\mathrm{div}, S_R))'$.

2. The exterior electromagnetic operator for Maxwell's equations. In this section we give an explicit construction of the electromagnetic operator T^ω defined for a tangential vector field $g^\omega \in TH^{-1/2}(\mathrm{curl}, S_R)$ by

$$T^\omega(g^\omega) = \mathcal{H}^\omega \wedge n, \tag{2.1}$$

where

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,\mathcal{H}^\omega = -i\omega\varepsilon_0\,\mathcal{E}^\omega \quad \text{in } R^3 \setminus \overline{B_R}, \\ \overrightarrow{\mathrm{curl}}\,\mathcal{E}^\omega = i\omega\mu_0\,\mathcal{H}^\omega \quad \text{in } R^3 \setminus \overline{B_R}, \\ \mathcal{E}^\omega_{S_R} = g^\omega \quad \text{on } S_R, \\ \lim_{r\to+\infty}\left(\sqrt{\mu_0}\,\mathcal{H}^\omega \wedge x - r\sqrt{\varepsilon_0}\,\mathcal{E}^\omega\right) = 0. \end{cases} \tag{2.2}$$

In order to prove the convergence of solutions to Maxwell's equations (1.1) as $\omega \to 0$, we also need the operator D^ω defined by

$$D^\omega : g^\omega \in TH^{-1/2}(\mathrm{curl}, S_R) \mapsto \mathcal{E}^\omega \cdot n. \tag{2.3}$$

Since $\mathcal{E}^\omega$ satisfies

$$\mathcal{E}^\omega = \frac{i}{\omega\varepsilon_0}\overrightarrow{\mathrm{curl}}\mathcal{H}^\omega,$$

it follows that

$$D^\omega(g^\omega) = \frac{i}{\omega\varepsilon_0}\mathrm{div}_{S_R} T^\omega(g^\omega),$$

and then the explicit expression of the operator D^ω will be deduced from that of the operator T^ω. Our basic aim in this section is to analyze the dependence of both the operators T^ω and D^ω on the frequency ω. It should be noted that a similar asymptotic analysis was considered by Lassas [23] from the point of view of inverse problems.

We start by representing the boundary data in terms of suitable vector basis functions on S_R. Following [13], let $(Y_l^m)_{-l\le m\le l}$ be an orthonormal sequence of spherical harmonics of order l on the unit sphere Σ, normalized such that

$$\int_\Sigma Y_l^m \cdot \overline{Y}_{l'}^{m'} = \delta_{l,l'}\,\delta_{m,m'}.$$

The basis functions for tangential fields on S_R are then

$$G_l^m = \frac{1}{\sqrt{l(l+1)}}\,\overrightarrow{\mathrm{grad}}_{S_R}\,Y_l^m \text{ and } R_l^m = n \wedge G_l^m \quad \text{for } -l \le m \le l,\ l \ge 1.$$

The tangential vector fields G_l^m and R_l^m are an orthonormal basis on the unit sphere (in the L^2 inner product). It follows that any tangential vector field $g \in TL^2(S_R)$ can be written in the form

$$g = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \alpha_l^m \, G_l^m + \beta_l^m \, R_l^m. \tag{2.4}$$

Spaces of either scalar or vector functions on S_R can be characterized by the summability of weighted sums of their expansion coefficients. Using the series coefficients (see [17]), the norm on the space $TH^s(S_R)$ can be characterized by

$$||g||^2_{TH^s(S_R)} = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \Big(1 + l(l+1)\Big)^s \Big(|\alpha_l^m|^2 + |\beta_l^m|^2\Big);$$

the norm on the space $H^s(S_R)$ by

$$||\varphi||^2_{H^s(S_R)} = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \Big(1 + l(l+1)\Big)^s |\varphi_l^m|^2,$$

where

$$\varphi = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \varphi_l^m \, Y_l^m;$$

the norm on the space $TH^{-1/2}(\text{curl}, S_R)$ by

$$||g||^2_{TH^{-1/2}(\text{curl},S_R)} = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \sqrt{1 + l(l+1)}\, |\beta_l^m|^2 + \frac{1}{\sqrt{1 + l(l+1)}}\, |\alpha_l^m|^2;$$

and the norm on the space $TH^{-1/2}(\text{div}, S_R)$ by

$$||g||^2_{TH^{-1/2}(\text{div},S_R)} = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \sqrt{1 + l(l+1)}\, |\alpha_l^m|^2 + \frac{1}{\sqrt{1 + l(l+1)}}\, |\beta_l^m|^2.$$

Throughout this paper, we assume for simplicity in exposition that $\varepsilon_0 \, \mu_0 = 1$ and $R = 1$. Now, if we expand the tangential vector field g^ω in the form

$$g^\omega = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \alpha_l^m(\omega) \, G_l^m + \beta_l^m(\omega) \, R_l^m, \tag{2.5}$$

we have the following explicit representation for $T^\omega(g^\omega)$ (see Appendix A):

$$T^\omega(g^\omega) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{\alpha_l^m(\omega)}{\gamma_l(\omega)} \, G_l^m + \frac{\beta_l^m(\omega)\gamma_l(\omega)}{i\omega} \, R_l^m, \tag{2.6}$$

where

$$\gamma_l(\omega) = 1 + \omega \frac{(h_l^{(1)})'(\omega)}{h_l^{(1)}(\omega)}. \tag{2.7}$$

From (2.6), we deduce that the operator D^ω introduced in (2.3) is given by

$$D^\omega(g^\omega) = \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{\sqrt{l(l+1)}}{\gamma_l(\omega)} \alpha_l^m(\omega) Y_l^m \tag{2.8}$$

for any $g^\omega \in TH^{-1/2}(\text{curl}, S_R)$ in the form (2.5).

Now, we summarize the mapping properties of the exterior electromagnetic operator T^ω.

LEMMA 2.1. *There exists a constant C such that for any $g \in TH^{-1/2}(\text{curl}, S_R)$ the following inequality holds:*

$$||T^\omega(g)||_{TH^{-1/2}(\text{div},S_R)} \leq C\, ||g||_{TH^{-1/2}(\text{curl},S_R)}. \tag{2.9}$$

Furthermore,

$$\Re e \Big(T^\omega(g), g\Big) > 0 \tag{2.10}$$

for any $g \neq 0$ in $TH^{-1/2}(\text{curl}, S_R)$, where $(,)$ denotes the duality between $TH^{-1/2}(\text{div}, S_R)$ and $TH^{-1/2}(\text{curl}, S_R)$.

Proof. (2.9) is proved in [17]. Thus, only the last part of the lemma needs proving. From

$$\Big(T^\omega(g), g\Big) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} i\omega \frac{|\alpha_l^m|^2}{\gamma_l(\omega)} + \frac{|\beta_l^m|^2 \gamma_l(\omega)}{i\omega},$$

we obtain that

$$\Re e \Big(T^\omega(g), g\Big) = \sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \omega \frac{|\alpha_l^m|^2}{|\gamma_l(\omega)|^2} \Im m\, (\gamma_l(\omega)) + \frac{|\beta_l^m|^2}{\omega} \Im m\, (\gamma_l(\omega)).$$

The fact that $\Im m\,(\gamma_l(\omega)) > 0$ (see, for instance, [27]) implies that (2.9) holds for any $g \neq 0$ in $TH^{-1/2}(\text{curl}, S_R)$. □

Our next result shows that D^ω is continuous as a map

$$D^\omega : TH^{-1/2}(\text{curl}, S_R) \mapsto H^{-1/2}(S_R).$$

LEMMA 2.2. *There exists a constant C such that for any $g \in TH^{-1/2}(\text{curl}, S_R)$ the following inequality holds:*

$$||D^\omega(g)||_{H^{-1/2}(S_R)} \leq C\, ||g||_{TH^{-1/2}(\text{curl},S_R)}. \tag{2.11}$$

Proof. Using the fact that there exist positive constants C_1 and C_2 such that $\forall\, l$ (see, for instance, [13])

$$C_1\, l \leq |\gamma_l(\omega)| \leq C_2\, l,$$

we obtain

$$\begin{aligned} ||D^\omega(g)||^2_{H^{-1/2}(S_R)} &= \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{l(l+1)}{|\gamma_l|^2(\omega)} \frac{1}{\sqrt{l(l+1)}} |\alpha_l^m|^2, \\ &\leq C \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{1}{\sqrt{l(l+1)}} |\alpha_l^m|^2, \\ &\leq C\, ||g||^2_{TH^{-1/2}(\text{curl},S_R)}, \end{aligned}$$

where the constant C is independent of g. The dependence of the constant C on the frequency ω can be explicitly characterized. ▯

Now we wish to analyze the asymptotic behavior of the operators T^ω and D^ω as the frequency ω tends to zero. We start by recalling some well-known results on the logarithmic derivative of the Hankel function $h_l^{(1)}$. Following [27], we have

$$\omega \frac{(h_l^{(1)})'(\omega)}{h_l^{(1)}(\omega)} = -\frac{p_l(\omega)}{q_l(\omega)} + i\frac{\omega}{q_l(\omega)},$$

where

$$\begin{cases} p_l(t) = c_l^0 + 2c_l^1 \dfrac{1}{t^2} + \cdots + (l+1)c_l^l \dfrac{1}{t^{2l}}, \\ q_l(t) = c_l^0 + c_l^1 \dfrac{1}{t^2} + \cdots + c_l^l \dfrac{1}{t^{2l}}, \end{cases}$$

and

$$c_l^m = \frac{(m+l)!(2m)!}{4^m (m!)^2 (l-m)!}.$$

Therefore, we can write $\gamma_l(\omega)$ in the form

$$\gamma_l(\omega) = 1 - \frac{(l+1)c_l^l + \cdots + (\omega)^{2l} c_l^0}{c_l^l + \cdots + (\omega)^{2l} c_l^0} + i\frac{(\omega)^{2l+1}}{c_l^l + \cdots + (\omega)^{2l} c_l^0}. \tag{2.12}$$

First, we have the following lemmas. The proofs are given in Appendix B.

Lemma 2.3. *The following inequality holds:*

$$\frac{(l+1-m)\, c_l^{l-m}}{c_l^l} < 1 \tag{2.13}$$

$\forall\, l \geq 1$ *and* $m = 1$ *to* l.

Lemma 2.4. *There exist* ω_0 *and a constant* C *independent of* ω *and* l *such that for* $0 < \omega < \omega_0$ *and* $l \geq 1$ *the following estimate holds:*

$$\left|\frac{1}{\gamma_l(\omega)} + \frac{1}{l}\right| \leq \frac{C}{l^2}\,\omega. \tag{2.14}$$

Next, from (2.12) and (2.13), we obtain the following results.

Lemma 2.5. *There exists* $\omega_0 > 0$ *such that*

$$\gamma_l(\omega) = \sum_{j=0}^{+\infty} \gamma_l^j \omega^j \quad \forall\, 0 < \omega < \omega_0$$

$\forall\, l \geq 1$ *(*ω_0 *is independent of* l*). Furthermore, for any* $j \geq 1$ *there exists a constant* C_j *independent of* l *such that*

$$\left|\frac{\gamma_l^j}{\gamma_l^0}\right| \leq C_j \quad \forall\, l \geq 1. \tag{2.15}$$

LEMMA 2.6. *Let $\left(\delta_l^j\right)_{j\geq 0, l\geq 1}$ be defined by*

$$\frac{1}{\gamma_l(\omega)} = \frac{1}{\gamma_l^0}\sum_{j=0}^{+\infty}\delta_l^j\omega^j. \tag{2.16}$$

For any $j \geq 1$ there exists a constant C_j independent of l such that

$$|\delta_l^j| \leq C_j \quad \forall\, l \geq 1. \tag{2.17}$$

With the help of these results, we obtain the following two results which are crucial for our asymptotic analysis at low frequencies.

LEMMA 2.7. *The operator*

$$D^\omega - D : TH^{-1/2}(\mathrm{curl}, S_R) \mapsto H^{-1/2}(S_R)$$

is compact, where D is defined for any $g \in TH^{-1/2}(\mathrm{curl}, S_R)$ in the form (2.4) *by*

$$D(g) = -\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{\frac{l+1}{l}}\,\alpha_l^m\,Y_l^m.$$

Furthermore, there exists ω_0 such that for $0 < \omega < \omega_0$ the following inequality holds:

$$||(D^\omega - D)(g)||_{H^{1/2}(S_R)} \leq C\,\omega\,||g||_{TH^{-1/2}(\mathrm{curl}, S_R)},$$

where the constant C independent of ω.

Proof. Let g be in $TH^{-1/2}(\mathrm{curl}, S_R)$. We have

$$(D^\omega - D)(g) = \sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{l(l+1)}\left(\frac{1}{\gamma_l(\omega)} + \frac{1}{l}\right)\alpha_l^m\,Y_l^m.$$

Therefore, by (2.14) we obtain that there exists a constant C independent of ω such that

$$\begin{aligned}||(D^\omega - D)(g)||^2_{H^{1/2}(S_R)} &\leq C\,\omega^2\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{1+l(l+1)}\,\frac{1}{l^2}\,|\alpha_l^m|^2\\ &\leq 3C\,\omega^2\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\frac{1}{\sqrt{1+l(l+1)}}\,|\alpha_l^m|^2 + \sqrt{1+l(l+1)}\,|\beta_l^m|^2\\ &\leq 3C\,\omega^2\,||g||^2_{TH^{-1/2}(\mathrm{curl}, S_R)},\end{aligned}$$

and thus we have the claim. □

The operator D is then in some sense the limit of the operator D^ω as ω tends to zero. Now, by using Lemma 2.5, we also verify without any difficulty that we have the following.

LEMMA 2.8. *There exists a constant C such that for any $g \in TH^{-1/2}(\mathrm{curl}, S_R)$ the following inequality holds:*

$$||T_j^\omega(g)||_{TH^{-1/2}(\mathrm{div}, S_R)} \leq C\,||g||_{TH^{-1/2}(\mathrm{curl}, S_R)}, \tag{2.18}$$

where the operator T_j^ω *is defined for any* $g \in TH^{-1/2}(\mathrm{curl}, S_R)$ *in the form* (2.4) *by*

$$\begin{aligned} T_j^\omega(g) &= \sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} i\omega \, \frac{\alpha_l^m}{\gamma_l^0} \, \sum_{p=0}^{j} \delta_l^p \omega^p \, G_l^m \\ &\qquad + \frac{\beta_l^m \gamma_l^0}{i\omega} \sum_{p=0}^{j} \frac{\gamma_l^p}{\gamma_l^0} \omega^p \, R_l^m . \end{aligned}$$

Proof. Let g be in $TH^{-1/2}(\mathrm{curl}, S_R)$. We have

$$\begin{aligned} ||T_j^\omega(g)||^2_{TH^{-1/2}(\mathrm{div}, S_R)} &= \frac{\varepsilon_0}{\mu_0} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \sqrt{1 + l(l+1)} \, \omega^2 \, \left| \frac{\alpha_l^m}{\gamma_l^0} \right|^2 \, \left| \sum_{p=0}^{j} \delta_l^p \omega^p \right|^2 \\ &\qquad + \frac{1}{\sqrt{1 + l(l+1)}} \, \left| \frac{\beta_l^m \gamma_l^0}{\omega} \right|^2 \left| \sum_{p=0}^{j} \frac{\gamma_l^p}{\gamma_l^0} \omega^p \right|^2 . \end{aligned}$$

Therefore, from Lemmas 2.5 and 2.6, we can show that there exists a constant C_j such that

$$\begin{aligned} ||T_j^\omega(g)||^2_{TH^{-1/2}(\mathrm{div}, S_R)} &\leq C_j \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{\sqrt{1 + l(l+1)}}{l^2} \, |\alpha_l^m|^2 + \frac{l^2}{\sqrt{1 + l(l+1)}} \, |\beta_l^m|^2 \\ &\leq 3C_j \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{1}{\sqrt{1 + l(l+1)}} \, |\alpha_l^m|^2 + \sqrt{1 + l(l+1)} \, |\beta_l^m|^2 \\ &\leq 3C_j \, ||g||^2_{TH^{-1/2}(\mathrm{curl}, S_R)}, \end{aligned}$$

and thus we have the claim. $\square$

3. The variational problem for Maxwell's equations. Several proofs of the existence of solutions to the scattering problem for nonhomogeneous Maxwell's equations in unbounded domains are now known. The proof of the existence and uniqueness of a solution for the case when ε and μ are in $\mathcal{C}^2$ was first given by Müller [25]. Abboud and Nédélec [2] proved existence and uniqueness for Maxwell's equations under the same regularity assumptions on the dielectric coefficients ε and μ but they are possibly discontinuous across the interface dielectric medium/exterior domain. They used the standard Sobolev space H^1 as a basis of their variational problem. Costabel [14] used another modified bilinear form coercive over H^1 to establish the existence of a solution in a bounded domain. Kirsch and Monk [17], [18] obtained a simple variational proof of the existence of a solution based on the use of the space $H(\mathrm{curl}, \Omega)$ where Ω is a smooth bounded domain. The crucial points in their analysis are the Hodge decomposition of the space $H(\mathrm{curl}, \Omega)$ and the compact embedding of the set $\{u \in H(\mathrm{curl}, \Omega) : \int_\Omega \varepsilon u \cdot \overrightarrow{\mathrm{grad}} \, \varphi = 0 \quad \forall \varphi \in H^1(\Omega)\}$ into $(L^2(\Omega))^3$. Note that the use of the Hodge decomposition of the space $H(\mathrm{curl}, \Omega)$ is a well-known idea for the study of Maxwell's equations in bounded domains (see Birman and Solomyak [5] and Leis [24]). Abboud [1] has analyzed the scattering of electromagnetic waves from periodic gratings in three dimensions. His method is very close to [17] but the compactness result that he used is quite different. More recently, Hazard and Lenoir [16] obtained by adding a regularizing term such as $\overrightarrow{\mathrm{grad}}\,\mathrm{div}$ in the time-harmonic Maxwell's equations an elliptic problem similar to the vector Helmholtz equation.

For the uniqueness of a solution to the scattering problem from inhomogeneities, several results are also known, but the list of such results is still incomplete. In fact, under certain regularity assumptions on ε and μ, namely ε and μ are $\mathcal{C}^{1,1}$ in every subdomain of the dielectric medium, it can be shown that a principle of unique continuation is valid (see Vogelsang [36], Colton and Kress [12]). As a consequence we would have that any solution of Maxwell's equations is unique. For our purposes, since ε and μ are assumed to be in $L^\infty(B_R)$, we have to suppose that the Maxwell system (1.1) has at most only one solution.

Now, following Kirsch and Monk [18], for instance, to obtain a variational formulation of the Maxwell system (1.1) in $H(\text{curl}, B_R)$ we shall first eliminate the magnetic field H^ω from (1.1). We simply substitute $H^\omega = \frac{1}{i\omega\mu}\overrightarrow{\text{curl}}\, E^\omega$ into $\overrightarrow{\text{curl}}\, H^\omega = -i\omega\varepsilon E^\omega$ to obtain

$$\overrightarrow{\text{curl}}\, \frac{1}{\mu}\, \overrightarrow{\text{curl}}\, E^\omega - \omega^2 \varepsilon\, E^\omega \;=\; 0 \quad \text{in } B_R. \tag{3.1}$$

If we multiply (3.1) by a test function Φ in $H(\text{curl}, B_R)$, integrate over B_R, and using integration by parts, we obtain

$$\int_{B_R} \frac{1}{\mu}\, \overrightarrow{\text{curl}}\, E^\omega \,\cdot\, \overrightarrow{\text{curl}}\, \Phi - \omega^2 \int_{B_R} \varepsilon\, E^\omega \,\cdot\, \Phi - \Big(\overrightarrow{\text{curl}}\, E^\omega \,\wedge\, n, \Phi_{S_R}\Big) \;=\; 0, \tag{3.2}$$

where the last term in this equation is to take, in the sense of the duality $TH^{-1/2}(\text{div}, S_R)$, $TH^{-1/2}(\text{curl}, S_R)$. Now, using the definition of the electromagnetic operator T^ω, we may write

$$\begin{aligned} &\int_{B_R} \frac{1}{\mu}\, \overrightarrow{\text{curl}}\, E^\omega \,\cdot\, \overrightarrow{\text{curl}}\, \Phi - \omega^2 \int_{B_R} \varepsilon\, E^\omega \,\cdot\, \Phi \\ &-i\omega\mu_0 \Big(T^\omega(E^\omega_{S_R}), \Phi_{S_R}\Big) \;=\; i\omega\mu_0 \,\Big(g^{in}_\omega, \Phi_{S_R}\Big), \end{aligned} \tag{3.3}$$

where

$$g^{in}_\omega \;=\; T^\omega(E^{in}_{\omega,S_R}) - H^{in}_\omega \,\wedge\, n. \tag{3.4}$$

We have from [17] or [27] the following.

LEMMA 3.1. *Suppose that the Maxwell system* (4.1) *admits at most one solution. Then the variational equation* (3.3) *is uniquely solvable in* $H(\text{curl}, B_R)$.

4. The derivation of the higher-order terms. This section is probably the most important section of this article because essentially all of what follows is either based on or motivated by the results we are about to discuss. We proceed formally to reduce the calculation of higher-order terms to the solutions of certain canonical problems. Then we shall show that each of these problems admits a unique solution. This will be accomplished with the help of the lemmas given above. The convergence of the electric and magnetic fields with these higher-order terms is the basic aim of the next section. Let us first recall that the electric and magnetic fields are solutions of the following scattering problem:

$$\left\{ \begin{aligned} &\overrightarrow{\text{curl}}\, H^\omega \;=\; -i\omega\varepsilon\, E^\omega \quad \text{in } B_R, \\ &\overrightarrow{\text{curl}}\, E^\omega \;=\; i\omega\mu\, H^\omega \quad \text{in } B_R, \\ &T^\omega(E^\omega_{S_R}) - H^\omega \,\wedge\, n \;=\; T^\omega(E^{in}_{\omega,S_R}) - H^{in}_\omega \,\wedge\, n \quad \text{on } S_R, \end{aligned} \right. \tag{4.1}$$

where we have replaced the Silver–Müller radiation condition by the exact nonlocal boundary condition

$$T^\omega(E^\omega_{S_R}) - H^\omega \wedge n = T^\omega(E^{in}_{\omega,S_R}) - H^{in}_\omega \wedge n \tag{4.2}$$

on the artificial boundary S_R.

Let us now formally expand the electric and the magnetic fields (E^ω, H^ω) in a power series of the frequency ω:

$$\begin{cases} E^\omega = \displaystyle\sum_{j=0}^{+\infty} E^j\,\omega^j, \\ H^\omega = \displaystyle\sum_{j=0}^{+\infty} H^j\,\omega^j, \end{cases} \tag{4.3}$$

where the fields (E^j, H^j) are in $H(\text{curl}, B_R) \times H(\text{curl}, B_R)$. Therefore, the tangential component of the electric field on S_R can be written in the form

$$E^\omega_{S_R} = \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} u^m_l(\omega)\,G^m_l + v^m_l(\omega)\,R^m_l,$$

where its expansion coefficients $u^{m,j}_l$ and $v^{m,j}_l$ have the following expansion forms with respect to the frequency ω:

$$\begin{cases} u^m_l(\omega) = \displaystyle\sum_{j=0}^{+\infty} u^{m,j}_l \omega^j, \\ v^m_l(\omega) = \displaystyle\sum_{j=0}^{+\infty} v^{m,j}_l \omega^j. \end{cases}$$

From the exact nonlocal boundary condition (4.2) on S_R we obtain that

$$\sqrt{\frac{\varepsilon_0}{\mu_0}}\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} i\omega\frac{1}{\gamma^0_l}\left(\sum_{j=0}^{+\infty}\delta^j_l\omega^j\right)\left(\sum_{j=0}^{+\infty}u^{m,j}_l\omega^j\right)G^m_l$$
$$+\frac{\gamma^0_l}{i\omega}\left(\sum_{j=0}^{+\infty}\frac{\gamma^j_l}{\gamma^0_l}\omega^j\right)\left(\sum_{j=0}^{+\infty}v^{m,j}_l\omega^j\right)R^m_l - \sum_{j=0}^{+\infty}\Big(H^j \wedge n\Big)\,\omega^j = g^{in}_\omega,$$

where g^{in}_ω is defined by (3.4). Now, using the fact that

$$\left(\sum_{j=0}^{+\infty}\delta^j_l\omega^j\right)\left(\sum_{j=0}^{+\infty}u^{m,j}_l\omega^j\right) = \sum_{j=0}^{+\infty}\left(\sum_{p=0}^{j}\delta^{j-p}u^{m,p}_l\right)\omega^j$$

and

$$\left(\sum_{j=0}^{+\infty}\frac{\gamma^j_l}{\gamma^0_l}\omega^j\right)\left(\sum_{j=0}^{+\infty}v^{m,j}_l\omega^j\right) = \sum_{j=0}^{+\infty}\left(\sum_{p=0}^{j}\frac{\gamma^{j-p}}{\gamma^0_l}v^{m,p}_l\right)\omega^j,$$

we obtain

$$\begin{aligned}&\sqrt{\frac{\varepsilon_0}{\mu_0}}\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} i\omega\frac{1}{\gamma_l^0}\sum_{j=0}^{+\infty}\left(\sum_{p=0}^{j}\delta_l^{j-p}u_l^{m,p}\right)\omega^j\, G_l^m\\ &+\frac{\gamma_l^0}{i\omega}\sum_{j=0}^{+\infty}\left(\sum_{p=0}^{j}\frac{\gamma_l^{j-p}}{\gamma_l^0}v_l^{m,p}\right)\omega^j\, R_l^m-\sum_{j=0}^{+\infty}\Big(H^j\,\wedge\,n\Big)\,\omega^j\;=\;g_\omega^{in}.\end{aligned}\tag{4.4}$$

If we assume that the incident electric field E_ω^{in} is analytic with respect to ω, we can write the quantity g_ω^{in} in the form

$$g_\omega^{in}\;=\;\frac{1}{\omega}\,g_{-1}^{in}+\sum_{j=0}^{+\infty}g_j^{in}\,\omega^j,$$

since $\gamma_l(\omega)$ is analytic with respect to ω in a real neighborhood of 0.

Now we shall show how to obtain the fields (E^j, H^j). Upon inserting the expansions (4.3) of (E^ω, H^ω) in the system (4.1), we get

$$\begin{cases}\overrightarrow{\operatorname{curl}}\,E^0\;=\;0\quad\text{in } B_R,\\ \overrightarrow{\operatorname{curl}}\,H^0\;=\;0\quad\text{in } B_R,\\ \operatorname{div}\varepsilon\,E^0\;=\;0\quad\text{in } B_R,\\ \operatorname{div}\mu\,H^0\;=\;0\quad\text{in } B_R\end{cases}\tag{4.5}$$

and more generally

$$\begin{cases}\overrightarrow{\operatorname{curl}}\,E^{j+1}\;=\;i\mu\,H^j\quad\text{in } B_R,\\ \overrightarrow{\operatorname{curl}}\,H^{j+1}\;=\;-i\varepsilon\,E^j\quad\text{in } B_R,\\ \operatorname{div}\varepsilon\,E^{j+1}\;=\;0\quad\text{in } B_R,\\ \operatorname{div}\mu\,H^{j+1}\;=\;0\quad\text{in } B_R.\end{cases}\tag{4.6}$$

To complete the derivation of the canonical problem satisfied by the field E^j or the field H^j, we need to find a boundary condition satisfied by these fields on S_R. Identifying different terms in the identity (4.4) and by following increasing powers of the frequency ω, we obtain

$$\begin{cases}-i\sqrt{\dfrac{\varepsilon_0}{\mu_0}}\displaystyle\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\gamma_l^0 v_l^{m,0}\,R_l^m\;=\;g_{-1}^{in},\\ -i\sqrt{\dfrac{\varepsilon_0}{\mu_0}}\displaystyle\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\gamma_l^0\left(\frac{\gamma_l^1}{\gamma_l^0}v_l^{m,0}+v_l^{m,1}\right)R_l^m-H^0\,\wedge\,n\;=\;g_0^{in}\end{cases}\tag{4.7}$$

and more generally

$$\begin{aligned}&i\sqrt{\frac{\varepsilon_0}{\mu_0}}\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\frac{1}{\gamma_l^0}\left(\sum_{p=0}^{j}\delta_l^{j-p}u_l^{m,p}\right)G_l^m\\ &-\gamma_l^0\left(\sum_{p=0}^{j+2}\frac{\gamma_l^{j+2-p}}{\gamma_l^0}v_l^{m,p}\right)R_l^m-H^{j+1}\,\wedge\,n\;=\;g_{j+1}^{in},\quad j\geq 0.\end{aligned}\tag{4.8}$$

We first derive a boundary condition on $S_R (R = 1)$ satisfied by the field E^0. Since

$$\left\{ \begin{array}{lcl} \operatorname{div}_{S_R} G_l^m & = & -\sqrt{l(l+1)}\, Y_l^m, \\ \operatorname{div}_{S_R} R_l^m & = & 0, \end{array} \right. \tag{4.9}$$

then by taking the surface divergence of the identity (4.4) for $j = 0$ we obtain

$$-\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{i}{\gamma_l^0} u_l^{m,0} \sqrt{l(l+1)}\, Y_l^m - \operatorname{div}_{S_R} \Big(H^1 \wedge n \Big) \;=\; \operatorname{div}_{S_R} \Big(g_1^{in} \Big). \tag{4.10}$$

Here we have used the fact that $\delta_l^0 = 1$. However, from (4.1) we see that

$$\operatorname{div}_{S_R} (H^1 \wedge n) \;=\; \overrightarrow{\operatorname{curl}}\, H^1 \, \cdot \, n \;=\; -i\varepsilon_0\, E^0 \, \cdot \, n \quad \text{on } S_R.$$

Thus

$$-\sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \frac{1}{\gamma_l^0} u_l^{m,0} \sqrt{l(l+1)}\, Y_l^m + \, E^0 \, \cdot \, n \;=\; -\frac{i}{\varepsilon_0} \operatorname{div}_{S_R} \Big(g_1^{in} \Big). \tag{4.11}$$

Now, let the operator

$$K : H^s(S_R) \mapsto H^{s+1}(S_R)$$

(for our purposes $s = -3/2$) be defined by

$$K(\varphi) \;=\; \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} -\frac{1}{\gamma_l^0}\, \varphi_l^m Y_l^m,$$

$(\gamma_l^0 = l)$ for $\varphi \in H^s(S_R)$ of the form $\varphi \;=\; \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \varphi_l^m Y_l^m$. Using the fact that $u_l^{m,0} \;=\; 0 \,\forall\, l \geq 1$ and $m = -l$ to l, the definition of the operator K, and the boundary condition on S_R (4.11), we obtain that E^0 satisfies on S_R

$$E^0 \, \cdot \, n - \, K(\operatorname{div}_{S_R} E_{S_R}^0) \;=\; -\frac{i}{\varepsilon_0} \operatorname{div}_{S_R} (g_1^{in}). \tag{4.12}$$

Let us observe that the operator K is related to the operator D by

$$D(g) = K(\operatorname{div}_{S_R} g) \quad \forall\, g \in TH^{-1/2}(\operatorname{curl}, S_R).$$

Now, to derive a boundary condition on S_R satisfied by the field H^0, we carry out the same procedure. Let us first observe that from (4.7) we can deduce that

$$\begin{aligned} \operatorname{div}_{S_R} (H_{S_R}^0) \quad = \quad & -i\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \sqrt{l(l+1)} \gamma_l^0\, v_l^{m,1}\, Y_l^m - \operatorname{curl}_{S_R} g_0^{in} \\ & -i\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \sqrt{l(l+1)} \gamma_l^1\, v_l^{m,0}\, Y_l^m. \end{aligned}$$

But

$$-\sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \sqrt{l(l+1)} v_l^{m,1}\, Y_l^m \;=\; \operatorname{div}_{S_R} (E^1 \wedge n) \;=\; i\mu_0\, H^0 \, \cdot \, n.$$

Thus,

$$i\mu_0\, H^0\ \cdot\ n + \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \sqrt{l(l+1)}v_l^{m,1}\, Y_l^m\ =\ 0. \tag{4.13}$$

Making use of the operator K, we can rewrite (4.13) in the more suitable form

$$\begin{aligned} H^0\ \cdot\ n -\ K(\mathrm{div}_{S_R}\, H^0_{S_R})\ &=\ -i\mu_0\, K(\mathrm{curl}_{S_R}\, g_0^{in}) \\ &- \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \frac{\sqrt{l(l+1)}\gamma_l^1}{\gamma_l^0}\, v_l^{m,0}\, Y_l^m. \end{aligned} \tag{4.14}$$

More generally, assume that the fields (E^i, H^i) are known for $i = 0$ to j; we wish to derive boundary conditions satisfied by the fields (E^{j+1}, H^{j+1}) on the fictive boundary S_R. From (4.8) it follows that

$$\begin{aligned} &\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} i\frac{1}{\gamma_l^0}\left(\sum_{p}^{j-1} \delta_l^{j-1-p} u_l^{m,p}\right) G_l^m \\ &-i\gamma_l^0 \left(\sum_{p=0}^{j+1} \frac{\gamma_l^{j+1-p}}{\gamma_l^0} v_l^{m,p}\right) R_l^m - H^j\ \wedge\ n\ =\ g_j^{in}, \end{aligned}$$

and so

$$\begin{aligned} -i\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \gamma_l^0 v_l^{m,j+1}\, R_l^m &= g_j^{in} + H^j\ \wedge\ n \\ &-i\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \frac{1}{\gamma_l^0}\left(\sum_{p=0}^{j-1} \delta_l^{j-1-p} u_l^{m,p}\right) G_l^m \\ &+i\sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \gamma_l^0 \left(\sum_{p=0}^{j} \frac{\gamma_l^{j+1-p}}{\gamma_l^0} v_l^{m,p}\right) R_l^m. \end{aligned} \tag{4.15}$$

Thus, the expansion coefficients $(v_l^{m,j+1})_{l\geq 1, -l\leq m\leq l}$ are determined. Furthermore, it is easy to see that

$$\begin{aligned} &\mathrm{div}_{S_R} \left\{\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} +i\frac{1}{\gamma_l^0}\left(\sum_{p=0}^{j+1} \delta_l^{j+1-p} u_l^{m,p}\right) G_l^m\right\} \\ &=\ \mathrm{div}_{S_R}\,(H^{j+2}\wedge n) + \mathrm{div}_{S_R}\,(g_{j+2}^{in}), \\ &=\ -i\varepsilon_0\, E^{j+1}\ \cdot\ n + \mathrm{div}_{S_R}\,(g_{j+2}^{in}). \end{aligned}$$

Thus,

$$i\varepsilon_0\, E^{j+1}\ \cdot\ n + i\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \frac{\sqrt{l(l+1)}}{\gamma_l^0}\left(\sum_{p=0}^{j+1} \delta_l^{j+1-p} u_l^{m,p}\right) Y_l^m\ =\ \mathrm{div}_{S_R}\,(g_{j+2}^{in}).$$

Finally, we obtain that E^{j+1} satisfies on S_R the following boundary condition:

$$\begin{aligned} &E^{j+1} \cdot n - K(\mathrm{div}_{S_R}(E^{j+1}_{S_R})) = -\frac{i}{\varepsilon_0}\mathrm{div}_{S_R}\left(g^{in}_{j+2}\right) \\ &+\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\frac{\sqrt{l(l+1)}}{\gamma_l^0}\left(\sum_{p=0}^{j}\delta_l^{j+1-p}u_l^{m,p}\right)Y_l^m. \end{aligned}$$

Similarly, the following identity holds for the field H^{j+1}:

$$\mathrm{div}_{S_R}(H^{j+1}_{S_R}) + i\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\gamma_l^0\sqrt{l(l+1)}\sum_{p=0}^{j+2}\frac{\gamma_l^{j+2-p}}{\gamma_l^0}v_l^{m,p}\,Y_l^m = \mathrm{curl}_{S_R}(g^{in}_{j+1}), \tag{4.16}$$

which can be rewritten in the form

$$\begin{aligned} &\mathrm{div}_{S_R}(H^{j+1}_{S_R}) + i\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\gamma_l^0\sqrt{l(l+1)}v_l^{m,j+2}\,Y_l^m = \mathrm{curl}_{S_R}(g^{in}_{j+1}) \\ &+i\sqrt{\frac{\varepsilon_0}{\mu_0}}\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\gamma_l^0\sqrt{l(l+1)}\sum_{p=0}^{j+1}\frac{\gamma_l^{j+2-p}}{\gamma_l^0}v_l^{m,p}\,Y_l^m, \end{aligned}$$

where the expansion coefficients $(v_l^{m,j+1})_{l\geq 1, -l\leq m\leq l}$ are known from the identity (4.15). Making use of the operator K, we see that

$$\begin{aligned} &K(\mathrm{div}_{S_R}(H^{j+1}_{S_R})) - i\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{l(l+1)}v_l^{m,j+2}\,Y_l^m = K(\mathrm{curl}_{S_R}(g^{in}_{j+1})) \\ &-i\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{l(l+1)}\sum_{p=0}^{j+1}\frac{\gamma_l^{j+2-p}}{\gamma_l^0}v_l^{m,p}\,Y_l^m. \end{aligned}$$

But

$$\begin{aligned} -\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{l(l+1)}v_l^{m,j+2}\,Y_l^m &= \mathrm{div}_{S_R}(E^{j+2}\wedge n) \\ &= i\mu_0\,H^{j+1}\cdot n. \end{aligned}$$

Thus

$$\begin{aligned} &K(\mathrm{div}_{S_R}(H^{j+1}_{S_R})) - H^{j+1}\cdot n = i\mu_0\,K(\mathrm{curl}_{S_R}(g^{in}_{j+1})) \\ &-\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\sqrt{l(l+1)}\sum_{p=0}^{j+1}\frac{\gamma_l^{j+2-p}}{\gamma_l^0}v_l^{m,p}\,Y_l^m \quad \text{on } S_R. \end{aligned}$$

To summarize, we have the following lemma.

LEMMA 4.1. *Assume that the electric and magnetic fields (E^ω, H^ω) admit the asymptotic expansions* (4.3). *Then we have*

$$\begin{cases} \overrightarrow{\mathrm{curl}}\,E^0 = 0 \quad \text{in } B_R, \\ \mathrm{div}\,\varepsilon\,E^0 = 0 \quad \text{in } B_R, \\ E^0\cdot n - K(\mathrm{div}_{S_R}(E^0_{S_R})) = -\frac{i}{\varepsilon_0}\mathrm{div}_{S_R}(g_1^{in}) \quad \text{on } S_R, \end{cases} \tag{4.17}$$

$$(4.18)\qquad \begin{cases} \overrightarrow{\mathrm{curl}}\, H^0 \;=\; 0 \quad \textit{in } B_R, \\ \mathrm{div}\, \mu\, H^0 \;=\; 0 \quad \textit{in } B_R, \\ H^0 \,\cdot\, n - K(\mathrm{div}_{S_R}\,(H^0_{S_R})) \;=\; -i\mu_0\, K(\mathrm{curl}_{S_R}\,(g_0^{in})) \\ -\displaystyle\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \sqrt{l(l+1)}\frac{\gamma_l^1}{\gamma_l^0} v_l^{m,0}\, Y_l^m \quad \textit{on } S_R, \end{cases}$$

$$(4.19)\qquad \begin{cases} \overrightarrow{\mathrm{curl}}\, E^{j+1} \;=\; i\mu\, H^j \quad \textit{in } B_R, \\ \mathrm{div}\, \varepsilon\, E^{j+1} \;=\; 0 \quad \textit{in } B_R, \\ E^{j+1} \,\cdot\, n - \; K(\mathrm{div}_{S_R}\,(E^{j+1}_{S_R})) \;=\; -\dfrac{i}{\varepsilon_0}\,\mathrm{div}_{S_R}\,(g_{j+2}^{in}) \\ +\displaystyle\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \frac{\sqrt{l(l+1)}}{\gamma_l^0}\left(\sum_{p=0}^{j}\delta_l^{j+1-p}u_l^{m,p}\right) Y_l^m \quad \textit{on } S_R, \end{cases}$$

and

$$(4.20)\qquad \begin{cases} \overrightarrow{\mathrm{curl}}\, H^{j+1} \;=\; -i\varepsilon\, E^j \quad \textit{in } B_R, \\ \mathrm{div}\, \mu\, H^{j+1} \;=\; 0 \quad \textit{in } B_R, \\ H^{j+1} \,\cdot\, n - K(\mathrm{div}_{S_R}\,(H^{j+1}_{S_R})) \;=\; -i\mu_0\, K(\mathrm{curl}_{S_R}\,(g_{j+1}^{in})) \\ -\displaystyle\sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \sqrt{l(l+1)}\sum_{p=0}^{j+1}\frac{\gamma_l^{j+2-p}}{\gamma_l^0} v_l^{m,p}\, Y_l^m \quad \textit{on } S_R. \end{cases}$$

Now, in order to prove the existence of a solution to the boundary-value problems (4.19) and (4.20) we must analyze in some detail the boundary conditions satisfied by the fields (E^j, H^j) on S_R. We shall prove the following lemma, which is essential to reduce the calculation of higher-order terms $(E^j, H^j)_{j\geq 1}$ to a certain canonical problem.

LEMMA 4.2. (a) *Assume that E^i is in $H(\mathrm{curl}, B_R)$ for $i = 1$ to j. Then the quantity*

$$-\frac{i}{\varepsilon_0}\mathrm{div}_{S_R}\,(g_{j+2}^{in}) + \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \frac{\sqrt{l(l+1)}}{\gamma_l^0}\left(\sum_{p=0}^{j}\delta_l^{j+1-p}u_l^{m,p}\right) Y_l^m$$

is in $H^{-1/2}(S_R)$.

(b) *Assume that E^i is in $H(\mathrm{curl}, B_R)$ for $i = 1$ to $j+1$. Then the quantity*

$$-i\mu_0\, K(\mathrm{curl}_{S_R}\,(g_{j+1}^{in})) - \sum_{l=1}^{+\infty}\sum_{m=-l}^{l} \sqrt{l(l+1)}\sum_{p=0}^{j+1}\frac{\gamma_l^{j+2-p}}{\gamma_l^0} v_l^{m,p}\, Y_l^m$$

is in $H^{-1/2}(S_R)$.

Proof. This lemma follows immediately from Lemma 2.8. □

We also need the following results.

Lemma 4.3. *The boundary-value problem*

$$\left\{\begin{array}{l} \operatorname{div} \varepsilon \,\overrightarrow{\operatorname{grad}}\, \psi \;=\; 0 \quad \textit{in } B_R, \\ \partial_n \psi -\; K(\Delta_{S_R}\, \psi) \;=\; f \in H^{-1/2}(S_R) \end{array}\right. \tag{4.21}$$

is uniquely solvable in $H^1(B_R)$ *within an additive constant.*

Proof. It is easy to see that if $\psi \in H^1(B_R)$ satisfies (4.21), then it is solution of the variational equation

$$a(\psi,\psi_t) = \int_{B_R} \varepsilon \,\overrightarrow{\operatorname{grad}}\, \psi \,\overrightarrow{\operatorname{grad}}\psi_t + \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} (l+1)\, \psi_l^m \,(\psi_t)_l^m \;=\; \int_{S_R} f\, \psi_t \quad \forall\, \psi_t \in H^1(B_R),$$

where ψ_l^m and $(\psi_t)_l^m$ are the expansion coefficents of ψ and ψ_t on S_R. Since

$$a\Big(\psi,\overline{\psi}\Big) \geq \int_{B_R} \varepsilon\, |\overrightarrow{\operatorname{grad}}\, \psi\,|^2,$$

the lemma follows then from the Lax–Milgram theorem by using a version of Poincaré's inequality. □

Next, we have to prove the following lemma.

Lemma 4.4. *Let* f *be in* $H(\operatorname{curl}, B_R)$ *such that* $\operatorname{div} f = 0$ *in* B_R*,* $q \in L^\infty(B_R)$*, and* $\varphi \in H^{-1/2}(S_R)$*. The boundary-value problem*

$$\left\{\begin{array}{l} \overrightarrow{\operatorname{curl}}\, u \;=\; f \quad \textit{in } B_R, \\ \operatorname{div} q\, u \;=\; 0 \quad \textit{in } B_R, \\ u \,\cdot\, n -\; K(\operatorname{div}_{S_R} u_{S_R}) \;=\; \varphi \quad \textit{on } S_R \end{array}\right. \tag{4.22}$$

is uniquely solvable in $H(\operatorname{curl}, B_R)$*.*

Proof. Since $f \in H(\operatorname{curl}, B_R)$ satisfies $\operatorname{div} f = 0$ in B_R, we can find $v \in H(\operatorname{curl}, B_R)$ such that $\operatorname{div} q\, v = 0$ and $f = \overrightarrow{\operatorname{curl}}\, v$ in B_R. Put $\tilde{u} = u - v$; then $\tilde{u}$ satisfies

$$\left\{\begin{array}{l} \overrightarrow{\operatorname{curl}}\, \tilde{u} \;=\; 0 \quad \text{in } B_R, \\ \operatorname{div} q\, \tilde{u} \;=\; 0 \quad \text{in } B_R, \\ \tilde{u} \,\cdot\, n -\; K(\operatorname{div}_{S_R} \tilde{u}_{S_R}) \;=\; \tilde{\varphi} \quad \text{on } S_R, \end{array}\right.$$

where

$$\tilde{\varphi} \;=\; \varphi - v \,\cdot\, n +\; K(\operatorname{div}_{S_R} v_{S_R})$$

is in $H^{-1/2}(S_R)$. But $\overrightarrow{\operatorname{curl}}\, \tilde{u} = 0$ in B_R implies that $\tilde{u} = \overrightarrow{\operatorname{grad}}\, \psi$, where ψ is in $H^1(B_R)$. Lemma 4.4 follows then immediately from Lemma 4.3. □

We note that the operator $K(\Delta_{S_R})$ is the Dirichlet–Neumann operator associated with the Laplace equation in $\mathbb{R}^3 \setminus \overline{B_R}$.

Finally, we can then state the following.

Theorem 4.5. *Assume that the electric and magnetic fields* (E^ω, H^ω) *admit the asymptotic expansions* (4.3)*. Then the fields* (E^j, H^j) *are determined as the unique solutions of the boundary-value problems* (4.17)–(4.20) *in* $H(\operatorname{curl}, B_R)$*.*

5. Convergence result. Now let M be an integer. Let (E^j, H^j) for $j = 0$ to M be defined recursively as the unique solutions of the boundary-value problems (4.17)–(4.20). Our aim in this section is to prove that the quantities

$$\left\| E^\omega - \sum_{j=0}^{M} E^j\,\omega^j \right\|_{(L^2(B_R))^3}, \qquad \left\| H^\omega - \sum_{j=0}^{M} H^j\,\omega^j \right\|_{(L^2(B_R))^3}$$

are of order $0(\omega^{M+1})$ which justifies the asymptotic expansions (4.3) of the electric and magnetic fields with respect to ω. The next theorem provides a new and simple proof of the convergence of the electric and magnetic fields solutions of scattering problems for Maxwell's equations as frequency tends to zero. In our proof of the convergence we shall need the following lemmas.

Lemma 5.1. *The boundary-value problem*

$$\begin{cases} \overrightarrow{\operatorname{curl}}\, v \;=\; 0 \textit{ in } B_R, \\ \operatorname{div} \varepsilon\, v \;=\; 0 \quad \textit{in } B_R, \\ v\,\cdot\, n -\, D(v_{S_R}) \;=\; 0 \quad \textit{on } S_R \end{cases} \tag{5.1}$$

has only the trivial solution in $H(\operatorname{curl}, B_R)$.

Proof. Since $D(v_{S_R}) = K(\operatorname{div}_{S_R} v_{S_R})$ the result follows from Lemma 4.4. □

Lemma 5.2. *Assume that* $q \in L^\infty(B_R)$ *and* $q = 1$ *in a neighborhood of* B_R. *Let*

$$V = \Big\{ w \in H(\operatorname{curl}, B_R), \operatorname{div} q\, w \;=\; 0 \text{ in } B_R, w\,\cdot\, n -\; D(w_{S_R}) \in H^{1/2}(S_R) \Big\}.$$

Then the embedding $V \hookrightarrow (L^2(B_R))^3$ *is compact.*

Proof. We first verify that the embedding

$$\Big\{ w \in H(\operatorname{curl}, B_R), \operatorname{div} q\, w \;=\; 0 \text{ in } B_R, w\,\cdot\, n -\; D(w_{S_R}) = 0 \Big\} \hookrightarrow (L^2(B_R))^3,$$

is compact. Let w be in this space. Since $\overrightarrow{\operatorname{curl}}\, w \in (L^2(B_R))^3$ there exists $\tilde{w}$ such that $\overrightarrow{\operatorname{curl}}\, \tilde{w} = \overrightarrow{\operatorname{curl}}\, w, \operatorname{div} qw = 0$, and $w\,.\,n = 0$ on S_R. Therefore, $w = \tilde{w} + \overrightarrow{\operatorname{grad}}\varphi$, where φ satisfies

$$\begin{cases} \operatorname{div} \varepsilon\, \overrightarrow{\operatorname{grad}}\, \psi \;=\; 0 \quad \text{in } B_R, \\ \partial_n \psi -\; K(\Delta_{S_R}\, \psi) \;\in H^{1/2}(S_R) \quad \text{on } S_R. \end{cases}$$

The boundary condition $\partial_n \psi - K(\Delta_{S_R}\, \psi) \in H^{1/2}(S_R)$ on S_R shows that $\varphi \in H^{3/2}(S_R)$ and so the claim is obtained from Weber's embedding theorems [37]. The lemma holds then from the fact that the embeddings $(H^1(B_R))^3 \hookrightarrow (L^2(B_R))^3$ and

$$\Big\{ w \in H(\operatorname{curl}, B_R), \operatorname{div} q\, w \;=\; 0 \text{ in } B_R, w\,\cdot\, n -\; D(w_{S_R}) = 0 \Big\} \hookrightarrow (L^2(B_R))^3$$

are compact. □

We can state the main result of this section.

Theorem 5.3. *Let* M *be an integer. Let* (E^j, H^j) *for* $j = 0$ *to* M *be defined recursively as the unique solutions of the boundary-value problems* (4.17)–(4.20). *There* *exists* $\omega_0 > 0$ *such that for* $0 < \omega < \omega_0$ *the following estimates hold:*

$$(5.2)\qquad \left\{ \begin{array}{ll} \left\| E^\omega - \sum_{j=0}^{M} E^j\,\omega^j \right\|_{(L^2(B_R))^3} & \leq C\,\omega^{M+1}, \\ \left\| H^\omega - \sum_{j=0}^{M} H^j\,\omega^j \right\|_{(L^2(B_R))^3} & \leq C\,\omega^{M+1}, \end{array} \right.$$

where the constant C *is independent of* ω.

Proof. It is easy to see that the field $E^\omega - E^0$ satisfies

$$(5.3)\qquad \left\{ \begin{array}{l} \overrightarrow{\mathrm{curl}}\,\frac{1}{\mu}\,\overrightarrow{\mathrm{curl}}\,(E^\omega - E^0) - \omega^2\varepsilon\,(E^\omega - E^0) \,=\, \omega^2\varepsilon E^0 \quad \text{in } B_R, \\ \mathrm{div}\,\varepsilon\,(E^\omega - E^0) \,=\, 0 \quad \text{in } B_R, \\ \overrightarrow{\mathrm{curl}}\,(E^\omega - E^0)\,\wedge\,n \,=\, i\omega\mu_0\left(T^\omega(E^\omega_{S_R} - E^0_{S_R}) - g^{in}_\omega\right) \\ +i\omega\mu_0\,T^\omega(E^0_{S_R}) \quad \text{on } S_R. \end{array} \right.$$

If we multiply the Maxwell's equations by $\overline{(E^\omega - E^0)}$, integrate over B_R, and use integration by parts, we obtain

$$(5.4)\qquad \begin{array}{l} \displaystyle\int_{B_R} \frac{1}{\mu}\,|\overrightarrow{\mathrm{curl}}\,(E^\omega - E^0)|^2 - \omega^2 \int_{B_R} \varepsilon\,|E^\omega - E^0|^2 \\ -i\omega\mu_0\Big(T^\omega(E^\omega_{S_R} - E^0_{S_R}), E^\omega_{S_R} - E^0_{S_R}\Big) \\ \displaystyle = \omega^2 \int_{B_R} \varepsilon\,E^0\,\cdot\,\overline{(E^\omega - E^0)} + i\omega\mu_0\left(g^{in}_\omega - T^\omega(E^0_{S_R}), E^\omega_{S_R} - E^0_{S_R}\right). \end{array}$$

Next, using the operator D^ω, we may write

$$(5.5)\qquad \begin{array}{rcl} (E^\omega - E^0)\,\cdot\,n & = & D^\omega(E^\omega_{S_R} - E^0_{S_R}) + D^\omega(E^0_{S_R}) + \,K(\mathrm{div}_{S_R}\,E^0_{S_R}) \\ & & \displaystyle -\frac{i}{\varepsilon_0}\,\mathrm{div}_{S_R}\,(g^{in}_1) + E^{in}_\omega\,\cdot\,n - D^\omega(E^{in}_{\omega,S_R}). \end{array}$$

We are now in position to prove Theorem 5.3. Let us first observe that

$$T^\omega(E^0_{S_R}) - \frac{1}{\omega}g^{in}_{-1} \,=\, \sqrt{\frac{\varepsilon_0}{\mu_0}}\sum_{l=1}^{+\infty}\sum_{m=-l}^{l}\frac{1}{\gamma^0_l}\,v^{m,0}_l\,G^m_l + 0(\omega),$$

where $0(\omega)$ is in $TH^{-1/2}(\mathrm{div}, S_R)$. If we assume that the quantity

$$\frac{1}{\omega}||E^\omega - E^0||_{(L^2(B_R))^3}$$

is not bounded then from (5.4) we can deduce that

$$v^\omega \,=\, \frac{E^\omega - E^0}{||E^\omega - E^0||_{(L^2(B_R))^3}}$$

is such that its curl is bounded (uniformly in ω) in $(L^2(B_R))^3$ and so it converges weakly in $\{w \in H(\mathrm{curl}, B_R), \mathrm{div}\varepsilon w = 0 \text{ in } B_R\}$ to a certain v. It is easy to see that $\overrightarrow{\mathrm{curl}} v = 0$, $\mathrm{div}\varepsilon v = 0$ in B_R and

$$v^\omega \cdot n - D(v^\omega_{S_R}) = (D^\omega - D)\left(\frac{E^\omega}{||E^\omega - E^0||_{(L^2(B_R))^3}}\right).$$

From Lemma 2.7, it follows that the following identity holds:

$$\left\|(D^\omega - D)\left(\frac{E^\omega}{||E^\omega - E^0||_{(L^2(B_R))^3}}\right)\right\|_{H^{1/2}(S_R)} \leq C\,\omega\,||v^\omega||_{H(\mathrm{curl},B_R)},$$

where C is a constant independent of ω and then by Lemma 5.1 we obtain that $v = 0$ and Lemma 5.2 implies that $v^\omega \to 0$ in $(L^2(B_R))^3$, strongly. This leads to a contradiction. The quantity $||E^\omega - E^0||_{(L^2(B_R))^3}$ is then of order ω. Now, in order to prove that $||H^\omega - H^0||_{(L^2(B_R))^3}$ is of order ω we observe that

$$\begin{aligned}
&\int_{B_R} \frac{1}{\mu}\,|\overrightarrow{\mathrm{curl}}\,(E^\omega - E^0 - \omega E^1)|^2 - \omega^2 \int_{B_R} \varepsilon\,|E^\omega - E^0 - \omega E^1|^2 \\
&-i\omega\mu_0\Big(T^\omega(E^\omega_{S_R} - E^0_{S_R} - \omega E^1_{S_R}), E^\omega_{S_R} - E^0_{S_R} - \omega E^1_{S_R}\Big) \\
&= \omega^2 \int_{B_R} \varepsilon\,E^0 \cdot \overline{(E^\omega - E^0 - \omega E^1)} \\
&+i\omega\mu_0\Big(g^{in}_\omega - T^\omega(E^0_{S_R}) - \omega T^\omega(E^1_{S_R}), E^\omega_{S_R} - E^0_{S_R} - \omega E^1_{S_R}\Big).
\end{aligned} \tag{5.6}$$

Since the quantity

$$\begin{aligned}
&|(g^{in}_\omega - T^\omega(E^0_{S_R}) - \omega T^\omega(E^1_{S_R}), E^\omega_{S_R} - E^0_{S_R} - \omega E^1_{S_R})| \\
&= 0(\omega)\,||E^\omega_{S_R} - E^0_{S_R} - \omega E^1_{S_R}||_{TH^{-1/2}(\mathrm{curl},S_R)},
\end{aligned}$$

it follows by the same argument that $||E^\omega - E^0 - \omega E^1||_{(L^2(B_R))^3}$ is of order ω^2. But

$$\frac{1}{\mu}\,\overrightarrow{\mathrm{curl}}\,(E^\omega - E^0 - \omega E^1) = i\omega(H^\omega - H^0).$$

Thus, from (5.6), we can deduce that $\omega^2\,||H^\omega - H^0||^2_{(L^2(B_R))^3}$ is of order ω^4 and then $||H^\omega - H^0||_{(L^2(B_R))^3}$ is of order ω. Now, since

$$\overrightarrow{\mathrm{curl}}\,\frac{1}{\mu}\,\overrightarrow{\mathrm{curl}}\left(E^\omega - \sum_{j=0}^{M} E^j\,\omega^j\right) - \omega^2\varepsilon\left(E^\omega - \sum_{j=0}^{M} E^j\,\omega^j\right) = \omega^{M+1}\varepsilon\Big(E^{M-1} + \omega E^M\Big)$$

in B_R, proceeding in a similar way, we obtain that the estimate

$$\left\|E^\omega - \sum_{j=0}^{M} E^j\,\omega^j\right\|_{(L^2(B_R))^3} = 0(\omega^{M+1})$$

holds for any integer M. To complete the proof of Theorem 5.3, we see that from

$$\left\|E^\omega - \sum_{j=0}^{M+1} E^j\,\omega^j\right\|_{(L^2(B_R))^3} = 0(\omega^{M+2}),$$

the variational equation

$$\begin{aligned}
&\int_{B_R} \frac{1}{\mu} \left| \overrightarrow{\mathrm{curl}} \left(E^\omega - \sum_{j=0}^{M+1} E^j\, \omega^j \right) \right|^2 - \omega^2 \int_{B_R} \varepsilon \left| E^\omega - \sum_{j=0}^{M+1} E^j\, \omega^j \right|^2 \\
&-i\omega\mu_0 \left(T^\omega \left(E^\omega_{S_R} - \sum_{j=0}^{M+1} E^j\, \omega^j \right), E^\omega_{S_R} - \sum_{j=0}^{M+1} E^j_{S_R}\, \omega^j \right) \\
&= \omega^{M+2} \int_{B_R} \varepsilon \left(E^M + \omega E^{M+1} \right) \cdot \left(\overline{E}^\omega - \sum_{j=0}^{M+1} \overline{E}^j\, \omega^j \right) \\
&+i\omega\mu_0 \left(g^{in}_\omega - T^\omega \left(\sum_{j=0}^{M+1} E^j_{S_R}\, \omega^j \right), E^\omega_{S_R} - \sum_{j=0}^{M+1} E^j_{S_R}\, \omega^j \right),
\end{aligned}$$

and the fact that the quantity

$$\left| \left(g^{in}_\omega - \sum_{j=0}^{M} T^\omega (E^j_{S_R}) \omega^j, E^\omega_{S_R} - \sum_{j=0}^{M} E^j_{S_R}\, \omega^j \right) \right|$$

is of order $0(\omega^M)\, ||E^\omega_{S_R} - \sum_{j=0}^{M} E^j_{S_R}\, \omega^j ||_{TH^{-1/2}(\mathrm{curl}, S_R)}$, we have

$$\omega^2 \int_{B_R} \mu \,|H^\omega - \sum_{j=0}^{M} H^j\, \omega^j|^2 \;=\; 0(\omega^{2M+4}).$$

Therefore, the quantity $||H^\omega - \sum_{j=0}^{M} H^j\, \omega^j||_{(L^2(B_R))^3}$ is of order $0(\omega^{M+1})$. □

Appendix A. The electromagnetic operator.

In this appendix, we give the explicit expression of the operator T^ω. Let us recall the standard vector basis functions for Maxwell's equations

$$\left\{ \begin{aligned} M_l^m(x) \;&=\; \overrightarrow{\mathrm{curl}} \Big(x\, h_l^{(1)}(\omega\sqrt{\varepsilon_0\mu_0})\, Y_l^m(x) \Big), \\ N_l^m(x) \;&=\; \frac{1}{i\omega\sqrt{\varepsilon_0\mu_0}}\, \overrightarrow{\mathrm{curl}}\, M_l^m(x), \end{aligned} \right.$$

where $h_l^{(1)}$ denotes the spherical Hankel functions of the first kind and order l. From the definitions of G_l^m and R_l^m, it is easy to see that

$$\begin{aligned} &-n \wedge \Big(n \,\wedge\, N_l^m(x) \Big) \\ &= -\frac{i}{\omega\sqrt{\varepsilon_0\mu_0}} \sqrt{l(l+1)} \left(h_l^{(1)}(\omega\sqrt{\varepsilon_0\mu_0}) + \omega\sqrt{\varepsilon_0\mu_0} \Big(h_l^{(1)} \Big)' (\omega\sqrt{\varepsilon_0\mu_0}) \right) G_l^m(x) \end{aligned}$$

and

$$-n \wedge \Big(n \,\wedge\, M_l^m(x) \Big) \;=\; -\sqrt{l(l+1)}\, h_l^{(1)}(\omega\sqrt{\varepsilon_0\mu_0})\, R_l^m(x)$$

on $S_R(R=1)$. Now, if we expand the tangential vector field g^ω in the form

$$g^\omega \;=\; \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} \alpha_l^m(\omega)\, G_l^m + \beta_l^m(\omega)\, R_l^m,$$

the solution of Maxwell's equations (2.2) satisfying the outgoing radiation condition can be written for $|x| > R = 1$ in the form (see, for instance, [13])

$$\mathcal{E}^\omega(x) \;=\; \sum_{l\geq 1} \sum_{m=-l}^{l} \Big(\delta_l^m(\omega)\, M_l^m(x) + \theta_l^m(\omega)\, N_l^m(x) \Big),$$

with uniform convergence on compact subsets of $|x| > R$, where

$$\delta_l^m(\omega) \;=\; -\frac{\beta_l^m(\omega)}{\sqrt{l(l+1)}\, h_l^{(1)}(\omega\sqrt{\varepsilon_0\mu_0})}$$

and

$$\theta_l^m(\omega) \;=\; \frac{i\omega\sqrt{\varepsilon_0\mu_0}\,\alpha_l^m(\omega)}{\sqrt{l(l+1)}\,\Big(h_l^{(1)}(\omega\sqrt{\varepsilon_0\mu_0}) + \omega\sqrt{\varepsilon_0\mu_0}\Big(h_l^{(1)}\Big)'(\omega\sqrt{\varepsilon_0\mu_0})\,\Big)}.$$

Since

$$\begin{cases} \overrightarrow{\text{curl}}\, M_l^m(x) \;=\; i\omega\sqrt{\varepsilon_0\mu_0}\, N_l^m(x), \\ \overrightarrow{\text{curl}}\, N_l^m(x) \;=\; -i\omega\sqrt{\varepsilon_0\mu_0}\, M_l^m(x), \end{cases}$$

the magnetic field $\mathcal{H}^\omega$ is given by

$$\mathcal{H}^\omega(x) \;=\; \frac{1}{\mu_0} \sum_{l\geq 1} \sum_{m=-l}^{l} \Big(\delta_l^m(\omega)\, N_l^m(x) - \theta_l^m(\omega)\, M_l^m(x) \Big).$$

Therefore, we have the following explicit representation for the operator T^ω:

$$T^\omega(g^\omega) \;=\; \sqrt{\frac{\varepsilon_0}{\mu_0}} \sum_{l=1}^{+\infty} \sum_{m=-l}^{l} i\omega \frac{\alpha_l^m(\omega)}{\gamma_l(\omega)}\, G_l^m + \frac{\beta_l^m(\omega)\gamma_l(\omega)}{i\omega}\, R_l^m,$$

where

$$\gamma_l(\omega) \;=\; 1 + \omega \frac{(h_l^{(1)})'(\omega)}{h_l^{(1)}(\omega)}.$$

Appendix B. Proofs of Lemmas 2.3–2.6.

In this appendix we shall give the proofs of Lemmas 2.3–2.6.

Proof of Lemma 2.3. It is easy to see that

$$\frac{l\, c_l^{l-1}}{c_l^l} \;<\; 1 \text{ and } \frac{c_l^m}{c_l^{m+1}} \;<\; 1, \quad m = 0 \text{ to } l-1.$$

By rewriting the quantity

$$\frac{(l+1-m)\,c_l^{l-m}}{c_l^l} = \frac{(l+1-m)\,c_l^{l-1}}{c_l^l}\,\frac{c_l^{l-2}}{c_l^{l-1}}\cdots\frac{c_l^{l-m}}{c_l^{l-m+1}},$$

we obtain the claim.

Proof of Lemma 2.4. For ω small enough, combining

$$\left|\frac{1}{\gamma_l(\omega)}+\frac{1}{l}\right| \le \omega \sup_{0<\omega_1<\omega}\left|\frac{\gamma_l'(\omega_1)}{\gamma_l^2(\omega_1)}\right|$$

and the following identity, which follows from the explicit form (2.12) of γ_l and Lemma 2.3, we obtain

$$\sup_{0<\omega_1<\omega}\left|\frac{\gamma_l'(\omega_1)}{\gamma_l^2(\omega_1)}\right| \le \frac{C}{l^2}\,\omega,$$

where $l \ge 1$ and the constant C is independent of ω and l, which thus gives the claim.

Next, from (2.12) and (2.13), we shall prove Lemma 2.5.

Proof of Lemma 2.5. Let

$$Q_l(t) = t^{2l}q_l(t) = c_l^l\left(1+\cdots+\frac{c_l^0}{c_l^l}t^{2l}\right).$$

From Lemma 2.3, it follows that there exists a complex neighborhood V of zero such that for any $l \ge 1$, $Q_l(t)$ does not have a zero in V. We can deduce that there exists a ball $B_r (r > 0)$ such that the function $\omega \mapsto \gamma_l(\omega)$ is analytic in B_r. (2.15) follows from the fact that there exists a constant C independent of ω such that

$$|\gamma_l(\omega)| \le C|\gamma_l^0| \quad \forall\, \omega \in \partial B_r.$$

Proof of Lemma 2.6. By using the fact that there exists a complex neighborhood V of zero such that

$$P_l(t) = t^{2l}p_l(t) + it^{2l}$$

does not have a zero in V, the identity (2.16) holds. (2.17) is easily deduced from (2.14).

REFERENCES

[1] T. Abboud, *Formulation variationnelle des équations de Maxwell dans un réseau bipériodique de R^3*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 245–248.

[2] T. Abboud and J. C. Nédélec, *Electromagnetic waves in an inhomogeneous medium,* J. Math. Anal. Appl., 164 (1992), pp. 40–58.

[3] H. Ammari, M. Laouadi, and J. C. Nédélec, *Low frequency behavior of solutions to electromagnetic scattering problems in chiral media,* SIAM J. Appl. Math., 58 (1998), pp. 1022–1042.

[4] H. Ammari and J. C. Nédélec, *Propagation d'ondes électromagnétiques à basses fréquences,* J. Math. Pures Appl., 77 (1998), pp. 839–849.

[5] M. Sh. Birman and M. Z. Solomyak, *L_2-Theory of the Maxwell operator in arbitrary domains,* Russian Math. Surveys, 42(6) (1987), pp. 75–96.

[6] O. P. Bruno and F. Reitich, *Solution of a boundary value problem for Helmholtz equation via variation of the boundary into the complex domain,* Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 317–340.

[7] O. P. Bruno and F. Reitich, *Numerical solution of diffraction problems: A method of variation of boundaries,* J. Opt. Soc. Amer. A, 10 (1993), pp. 1168–1175.

[8] O. P. Bruno and F. Reitich, *Numerical solution of diffraction problems: A method of variation of boundaries. Part* II. *Finitely conducting gratings, Padé approximants, and singularities,* J. Opt. Soc. Amer. A, 10 (1993), pp. 2307–2316.

[9] O. P. Bruno and F. Reitich, *Numerical solution of diffraction problems: A method of variation of boundaries. Part* III. *Doubly periodic gratings,* J. Opt. Soc. Amer. A, 10 (1993), pp. 2551–2562.

[10] O. P. Bruno and F. Reitich, *Approximation of analytic functions: A method of enhanced convergence,* Math. Comp., 63 (1994), pp. 195–214.

[11] O. P. Bruno and F. Reitich, *Calculation of electromagnetic scattering via boundary variations and analytic continuation,* ACES J., 11 (1996), pp. 17–31.

[12] D. Colton and R. Kress, *Time harmonic electromagnetic waves in an inhomogeneous medium,* Proc. Roy. Soc. Edinburgh Sect. A, 116 (1990), pp. 279–293.

[13] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory,* Springer-Verlag, Berlin, 1992.

[14] M. Costabel, *A coercive bilinear form for Maxwell's equations,* J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[15] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 3, *Spectral Theory and Applications,* Springer-Verlag, Berlin, 1990.

[16] C. Hazard and M. Lenoir, *On the solution of time-harmonic scattering problems for Maxwell's equations,* SIAM J. Math. Anal., 27 (1996), pp. 1597–1630.

[17] A. Kirsch and P. Monk, *A finite element/spectral method for approximating the time-harmonic Maxwell system in R^3,* SIAM J. Appl. Math., 55 (1995), pp. 1324–1344.

[18] A. Kirsch and P. Monk, *Corrigendum to "A finite element/spectral method for approximating the time-harmonic Maxwell system in R^3,* SIAM J. Appl. Math., 55 (1995), pp. 1324–1344," SIAM J. Appl. Math., 58 (1998), pp. 2024–2028.

[19] R. E. Kleinman, *Low frequency electromagnetic scattering,* in Electromagnetic Scattering, P. L. E. Uslenghi, ed., Academic Press, New York, 1978, pp. 1–28.

[20] R. E. Kleinman and T. B. A. Senior, *Rayleigh scattering*, in Low and High Frequency Asymptotics, V. K. Varadan and V. V. Varadan, eds., North-Holland, Amsterdam, 1986, pp. 1–70.

[21] R. Kress, *On the limiting behaviour of solutions to boundary integral equations associated with time harmonic wave equations for small frequencies,* Math. Methods Appl. Sci., 1 (1979), pp. 89–100.

[22] G. A. Kriegsmann and E. L. Reiss, *Low frequency scattering by local inhomogeneities,* SIAM J. Appl. Math., 43 (1983), pp. 923–934.

[23] M. Lassas, *The impedance imaging as a low frequency limit,* Inverse Problems, 13 (1997), pp. 1503–1518.

[24] R. Leis, *Initial Boundary Value Problems in Mathematical Physics,* Teubner and Wiley, Stuttgart, 1986.

[25] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves,* Springer-Verlag, Berlin, 1969.

[26] C. Müller and H. Niemeyer, *Greensche tensoren and asymptotische gesetze der elektromagnetischen hohlraumschwingungen,* Arch. Rational Mech. Anal., 7 (1961), pp. 305–348.

[27] J.-C. Nédélec, *Ondes acoustiques et électromagnétiques, Équations Integrales,* École Polytechnique, 1996.

[28] L. Paquet, *Problèmes mixtes pour le système de Maxwell,* Ann. Fac. Sci. Toulouse Math., 4 (1982), pp. 103–141.

[29] R. Picard, *On the low frequency asymptotics in electromagnetic theory,* J. Reine Angew. Math., 354 (1984), pp. 50–73.

[30] A. G. Ramm, *Iterative Methods for Calculating Static Fields and Wave Scattering by Small Bodies,* Springer-Verlag, Berlin, 1982.

[31] A. G. Ramm and E. Somersalo, *Electromagnetic inverse problems with surface measurements at low frequencies,* Inverse Problems, 5 (1989), pp. 1107–1116.

[32] A. G. Ramm, O. L. Weaver, N. Weck, and K. J. Witsch, *Dissipative Maxwell's equations at low frequencies*, Math. Methods Appl. Sci., 13 (1990), pp. 305–322.

[33] Lord Rayleigh, *On the electromagnetic theory of light,* Phil. Mag., 12 (1881), pp. 81–101.

[34] A. Stevenson, *Solution of electromagnetic scattering problems as power series in the ratio (dimension of scatter/wave length),* J. Appl. Phys., 24 (1953), pp. 1134–1142.

[35] M. E. Taylor, *Partial Differential Equations* II, Appl. Math. Sci. 116, Springer-Verlag, Berlin, 1996.

[36] V. Vogelsang, *On the strong unique continuation principle for inequalities of Maxwell type,* Math. Ann., 289 (1991), pp. 285–295.

[37] Ch. Weber, *A local compactness theorem for Maxwell's equations,* Math. Methods Appl. Sci., 2 (1980), pp. 12–25.

[38] N. Weck and K. J. Witsch, *Low-frequency asymptotics for dissipative Maxwell's equations in bounded domains,* Math. Methods Appl. Sci., 13 (1990), pp. 81–93.

[39] P. Werner, *Randwertprobleme für die zeitabhängigen Maxwellschen gleichungen mit variablen koeffizienten,* Arch. Rational Mech. Anal., 18 (1965), pp. 167–195.

[40] P. Werner, *On the behaviour of stationary electromagnetic wave fields for small frequencies*, J. Math. Anal. Appl., 15 (1966), pp. 447–496.

[41] P. Werner, *Über das verhalten elektromagnetischer felder für kleine frequenzen in mehrfach zusammenhängenden gebieten* I, J. Reine Angew. Math., 278 (1975), pp. 365–397.

[42] P. Werner, *Über das verhalten elektromagnetischer felder für kleine frequenzen in mehrfach zusammenhängenden gebieten* II, J. Reine Angew. Math., 280 (1976), pp. 98–121.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 862–893

LOCALIZATION OF SOLUTIONS OF EXTERIOR DOMAIN PROBLEMS FOR THE POROUS MEDIA EQUATION WITH RADIAL SYMMETRY*

B. H. GILDING† AND J. GONCERZEWICZ‡

Abstract. The paper concerns the radially symmetric Cauchy–Dirichlet and Cauchy–Neumann problems for the porous media equation in the domain comprising the spatial variable and the temporal variable t in the exterior of the unit ball in $\mathbb{R}^n$ and the bounded interval $(0, T)$, respectively. The subject of study is the behavior of solutions when the initial data are compactly supported and the boundary data become unbounded as $t \uparrow T$. Necessary and sufficient conditions for localization, estimates of the size of the blow-up set, and a number of allied results are obtained.

Key words. porous media equation, radial symmetry, localization, blow-up, peaking, comparison principle, self-similar solution

AMS subject classifications. 35K65, 35K55, 35B99, 35R35

PII. S0036141098344506

1. Statement of problems and results.

We consider two problems for the nonlinear partial differential equation

$$\frac{\partial u}{\partial t} = r^{1-n} \frac{\partial}{\partial r} \left(r^{n-1} \frac{\partial u^m}{\partial r} \right), \tag{1.1}$$

where

$$n \geq 1 \quad \text{and} \quad m > 1$$

are real parameters, for (r, t) in the domain

$$Q := (1, \infty) \times (0, T) \quad \text{with} \quad 0 < T < \infty.$$

Both problems have the initial condition

$$u(r, 0) = u_0(r) \quad \text{for } 1 < r < \infty, \tag{1.2}$$

where u_0 is a bounded, nonnegative, measurable function defined on the interval $(1, \infty)$ which vanishes for large r.

Problem 1.1. Equation (1.1) in Q subject to (1.2) and the Dirichlet boundary condition

$$u(1, t) = f(t) \quad \text{for } 0 < t < T.$$

*Received by the editors September 11, 1998; accepted for publication (in revised form) September 23, 1999; published electronically April 4, 2000.

http://www.siam.org/journals/sima/31-4/34450.html

†Faculty of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands (b.h.gilding@math.utwente.nl).

‡Institute of Mathematics, University of Wrocław, pl. Grunwaldzki 2/4, 50 - 384 Wrocław, Poland (goncerz@math.uni.wroc.pl). This author was supported by KBN grants 1066/P3/93/04 and 481/P03/97/12.

Problem 1.2. Equation (1.1) in Q subject to (1.2) and the Neumann boundary condition

$$-\frac{\partial u^m}{\partial r}(1,t) = g(t) \quad \text{for } 0 < t < T.$$

In these problems f and g are nonnegative, measurable functions defined on the interval $(0,T)$ which are bounded on every interval $(0,\tau)$ with $0 < \tau < T$.

Problems 1.1 and 1.2 are prompted by the study of the porous media equation $\partial u/\partial t = \Delta u^m$ in which t denotes time and Δ the Laplace operator in $\mathbb{R}^n$. This equation is well known as an outstanding benchmark for the study of quasilinear degenerate parabolic equations of second order. In contrast to a nondegenerate parabolic equation, when $m > 1$ this equation displays finite speed of propagation. This is to say that in an unbounded spatial domain with initial data possessing compact support, the solution retains compact support with respect to the spatial variable at all later times. The boundary of the support of the solution constitutes an interface with some established regularity. Simultaneously the solution may fail to solve the equation classically at this free boundary. It is known that the support of the solution cannot shrink in time and in an infinite time-span will eventually fill the whole spatial domain [3, 4, 23, 27, 36]. The present interest is in the solution of the porous media equation in the spatial domain comprising the exterior of the unit ball in $\mathbb{R}^n$ with Dirichlet or Neumann boundary conditions on the surface of the ball. Specifically, the interest is in the behavior of the solution and its support when the boundary data become unbounded as some critical finite time T is approached. A pertinent question, in this case, is whether or not the support of the solution may fill the whole domain at the critical time. If it does not, and the support of the solution remains bounded, the solution is said to exhibit localization. A closely related question is whether or not the volume of the subdomain in which the solution itself becomes unbounded as the critical time is approached is bounded. If this volume is bounded, the solution is said to display effective localization [31]. As a foundation for analyzing these questions we shall consider the restriction to radially symmetric solutions of the problems. In this event one arrives at Problems 1.1 and 1.2, with r denoting radius and n a positive integer.

The last thirty years have seen considerable advances in the understanding of the phenomena of localization and effective localization of solutions of initial-boundary-value problems for degenerate and nondegenerate quasilinear parabolic equations of second order (which include the porous media equation as a special case) and in domains of one or more spatial dimensions with diverse geometries. In particular, much knowledge has been acquired from the investigation of self-similar and "approximately self-similar" solutions of the equation in hand. For an encyclopedic and instructive overview of the study of the phenomena of localization and effective localization, including an exhaustive survey of antecedent literature, we recommend to the reader the monograph [31].

Notwithstanding the above remarks, excepting in recent papers of Shishkov [32] and Shishkov and Shchelkov [33], the questions of localization for Problems 1.1 and 1.2 for $n > 1$ do not appear to have been treated previously in the literature. Analysis has concentrated on the Cauchy–Dirichlet problem in rectangular spatial domains with convenient conditions on the boundary data, and especially on the Cauchy–Dirichlet problem in the case $n = 1$ under the assumption that the boundary data function f

is continuous and increasing. In this context it has been said that solutions are in a *peaking regime* and T is called the *peaking time* [15, 30].

To clarify our study further, let us state what is understood by a solution of Problems 1.1 and 1.2.

Definition 1.3. *A real function u defined on Q is said to be a weak solution of Problem* 1.1 *if, for any $0 < \tau < T$, u is nonnegative, bounded, and continuous in $(1,\infty) \times (0,\tau]$ and u satisfies the identity*

$$\int_0^\tau \int_1^\infty \left\{ r^{n-1} u \frac{\partial \varphi}{\partial t} + u^m \frac{\partial}{\partial r} \left(r^{n-1} \frac{\partial \varphi}{\partial r} \right) \right\} dr\, dt \\ + \int_1^\infty r^{n-1} u_0(r) \varphi(r,0)\, dr + \int_0^\tau f^m(t) \frac{\partial \varphi}{\partial r}(1,t)\, dt = 0$$

for all nonnegative $\varphi \in C^{2,1}([1,\infty) \times [0,\tau])$ which vanish for $t = \tau$, $r = 1$, and large r.

Definition 1.4. *A real function u defined on Q is said to be a weak solution of Problem* 1.2 *if, for any $0 < \tau < T$, u is nonnegative, bounded, and continuous in $(1,\infty) \times (0,\tau]$ and u satisfies the identity*

$$\int_0^\tau \int_1^\infty \left\{ r^{n-1} u \frac{\partial \varphi}{\partial t} + u^m \frac{\partial}{\partial r} \left(r^{n-1} \frac{\partial \varphi}{\partial r} \right) \right\} dr\, dt \\ + \int_1^\infty r^{n-1} u_0(r) \varphi(r,0)\, dr + \int_0^\tau g(t) \varphi(1,t)\, dt = 0$$

for all nonnegative $\varphi \in C^{2,1}([1,\infty) \times [0,\tau])$ which vanish for $t = \tau$ and large r and are such that $\partial\varphi/\partial r$ vanishes for $r = 1$.

The theory developed for initial-boundary-value problems for the porous media equation in one spatial dimension [3, 4, 23, 27] readily extends to Problems 1.1 and 1.2. This can be explained formally by the observation that (1.1) can be written as $\partial u/\partial t = \partial^2 u^m/\partial r^2 + (n-1)r^{-1}\partial u^m/\partial r$, which differs from the porous media equation in one spatial dimension merely by an extra term containing a first order derivative preceded by a coefficient of fixed sign which is not singular for $r > 1$. Correspondingly, Problems 1.1 and 1.2 each admit a unique weak solution u which has a derivative $\partial u^m/\partial r$ in $C(Q)$, is such that $(\partial u^m/\partial r)(r,\cdot) \in L^1(0,t)$ for every $(r,t) \in Q$, and is a classical solution of (1.1) at any point in Q where it is positive. For details, see [19].

Set

$$\zeta_0 := \sup \left\{ r \in (1,\infty) : \int_r^\infty u_0(x)\, dx > 0 \right\},$$

with the convention that this is equal to 1 if the supremum is taken over the empty set. Since u_0 vanishes for large r, this quantity is finite. Furthermore, by results in [18], if one defines

$$\zeta(t) := \sup\{ r \in (1,\infty) : u(r,t) > 0 \} \quad \text{for } 0 < t < T$$

with the convention that this variable is equal to 1 if the supremum is taken over the empty set, then ζ is a nondecreasing continuous function on $(0,T)$ such that $\zeta(t) \to \zeta_0$ as $t \downarrow 0$. Thus one can define

$$\zeta(T) := \sup\{\zeta(t) : 0 < t < T\}. \tag{1.3}$$

Following [11, 15, 31], u will be said to be *localized* if $\zeta(T) < \infty$.

For a weak solution u of Problem 1.1 or 1.2 we define the *blow-up set*

$$\Omega := \left\{ r \in (1,\infty) : \limsup_{t \uparrow T} u(r,t) = \infty \right\}. \tag{1.4}$$

In applications where the porous media equation is considered as a nonlinear model of heat conduction, this blow-up set is sometimes referred to as the *zone of heat intensification*. An alternative definition for such a set is

$$\Omega^* := \left\{ r \in (1,\infty) : \limsup_{t \uparrow T} \int_r^\infty x^{n-1} u(x,t)\, dx = \infty \right\}. \tag{1.5}$$

Considering (1.1) once more as a model of nonlinear heat conduction, this set represents the complement of the region in which the total energy remains finite. Under our standing convention that the supremum of the empty set is 1, let

$$\omega := \sup \Omega \quad \text{and} \quad \omega^* := \sup \Omega^*. \tag{1.6}$$

In view of a comparison principle for solutions of Problems 1.1 and 1.2, which is the natural extension of a similar principle for solutions of the porous media equation in one spatial dimension (details of which can be found in [19]), both of the blow-up sets Ω and Ω^* are either empty or an interval with infimum 1, i.e.,

$$(1,\omega) \subseteq \Omega \subseteq \overline{(1,\omega)} \quad \text{and} \quad (1,\omega^*) \subseteq \Omega^* \subseteq \overline{(1,\omega^*)}.$$

Furthermore, u and $\partial u^m / \partial r$ are continuous in $(\omega,\infty) \times (0,T]$,

$$\zeta(T) < \infty \quad \text{if and only if} \quad \omega < \infty,$$

and

$$\Omega^* \subseteq \Omega \subseteq (1, \zeta(T)). \tag{1.7}$$

Following [30, 31], it will be said that u is *effectively localized* if $\omega < \infty$, and *metastably localized* if $\omega^* < \infty$.

In the case $n = 1$, a definitive characterization of the occurrence of localization under the assumption that u_0 is continuous on $[1,\infty)$ and that f is continuous and monotonic increasing on $[0,T)$ and satisfies the compatibility condition $f(0) = u_0(1)$ was published in [15]. It was proven that localization occurs if and only if

$$\mathcal{C}^* := \limsup_{t \uparrow T} \frac{\int_0^t f^m(s)\, ds}{f(t)} \tag{1.8}$$

is finite. Moreover, defining

$$\ell(r) := \limsup_{t \uparrow T} (T-t)^{1/(m-1)}\, u(r,t) \quad \text{for } r > 1 \tag{1.9}$$

for a weak solution u of Problem 1.1, if localization occurs then $\ell(r) < \infty$ for all $r > 1$, whereas if localization does not occur then $\ell(r) = \infty$ for all $r > 1$. Furthermore, it was shown that if $r \in \Omega$ then $u(r,t) \to \infty$ as $t \uparrow T$. Finally in [15], estimates of the size of the blow-up set Ω were derived.

Shortly after the article [15] was published, another paper [11] in which localization for Problem 1.1 with $n = 1$ was characterized appeared. This article, by Cortazar and Elgueta, examined the problem under the assumption that u_0 is continuous on $(1, \infty)$ and that f is continuous on $(0, T)$. The result established was that Problem 1.1 exhibits localization if and only if

$$\mathcal{C} := \limsup_{t \uparrow T} (T - t)^{1/(m-1)} \int_0^t f^m(s)\, ds \tag{1.10}$$

is finite. It turns out that this criterion of Cortazar and Elgueta [11] represents the superior result, since it does not rely on the assumption that f is monotonic, and, as we shall see in an appendix, if f is continuous and nondecreasing, the conditions $\mathcal{C} < \infty$ and $\mathcal{C}^* < \infty$ are equivalent. The Cauchy–Neumann problem with $n = 1$ was also examined by Cortazar and Elgueta, under the assumption that u_0 is Lipschitz continuous on $[1, \infty)$. For this problem, they showed that localization occurs if and only if

$$\mathcal{C} := \limsup_{t \uparrow T} (T - t)^{1/(m-1)} \int_0^t g(s)\, ds \tag{1.11}$$

is finite.

The first contribution of the present paper will be to demonstrate the following.

THEOREM 1.5. *For any weak solution u of Problem* 1.1 *or Problem* 1.2 *there holds $\omega^* = \omega$.*

It follows that for a solution of Problem 1.1 or 1.2 the concepts of localization, effective localization, and metastable localization are all equivalent. Moreover, with the possible exception $\Omega^* = (1, \omega)$ and $\Omega = (1, \omega]$ with $1 < \omega < \infty$, the blow-up sets Ω and Ω^* coincide.

Specifically concerning the Cauchy–Dirichlet problem, we prove the next result.

THEOREM 1.6. *Let u be a weak solution of Problem* 1.1 *and let $\mathcal{C}$ be defined by* (1.10)*. Then $\omega < \infty$ if and only if $\mathcal{C} < \infty$.*

Thus the result of Cortazar and Elgueta [11] can be obtained under weaker assumptions on the initial and boundary data. Moreover, this result extends from $n = 1$ to every $n \geq 1$.

We shall also extend the results on the nature of blow-up which were obtained in [15] by demonstrating the following two theorems. The extension represents a weakening of the assumptions on the initial and boundary data, an enlargement on the conclusions, and, last but not least, a generalization from $n = 1$ to $n \geq 1$. By the statement that f is nondecreasing, we mean that $\operatorname{ess\,sup}\{f(s) : 0 < s < t\} \leq \operatorname{ess\,inf}\{f(s) : t < s < T\}$ for all $0 < t < T$.

THEOREM 1.7. *Let u be a weak solution of Problem* 1.1 *and $\mathcal{C}$ be specified by* (1.10)*. Define $\ell(r)$ by* (1.9) *and*

$$L(r) := \begin{cases} \mathcal{C} & \text{for } r = 1, \\ \limsup_{t \uparrow T} (T - t)^{1/(m-1)} \int_0^t u^m(r, s)\, ds & \text{for } r > 1. \end{cases} \tag{1.12}$$

(i) *Suppose that $\mathcal{C} < \infty$. Then $\ell^{m-1}, L \in C(1, \infty) \cap W^{1,\infty}_{\mathrm{loc}}(1, \infty)$, ℓ and L are strictly decreasing on $(1, \omega]$ and $[1, \omega]$, respectively, and $\ell \equiv L \equiv 0$ on (ω, ∞).*

(ii) *Suppose that $\mathcal{C} = \infty$. Then $\ell(r) = L(r) = \infty$ for all $r > 1$.*

Furthermore, if f is nondecreasing, then $u(\cdot, t) \to \infty$ as $t \uparrow T$ uniformly on $(1, r)$ for every $1 < r < \omega$.

THEOREM 1.8. *Let u_1 and u_2 be two weak solutions of Problem* 1.1 *with corresponding initial data functions $u_{0,i}$, boundary data functions f_i, parameters $\mathcal{C}_i$ defined by* (1.10)*, and functions ℓ_i and L_i defined by* (1.9) *and* (1.12) *for $i = 1, 2$. If*

$$\limsup_{t\uparrow T} (T-t)^{1/(m-1)} \int_0^t \{f_1^m(s) - f_2^m(s)\}\, ds \le 0, \tag{1.13}$$

then $L_1(r) \le L_2(r)$ for all $r \ge 1$. Moreover, if

$$(T-t)^{1/(m-1)} \int_0^t \max\{f_1^m(s) - f_2^m(s), 0\}\, ds \to 0 \qquad \text{as } t \uparrow T, \tag{1.14}$$

then $\ell_1(r) \le \ell_2(r)$ for all $r > 1$.

Regarding the size of the blow-up sets, we establish the next estimate.

THEOREM 1.9. *Let u be a weak solution of Problem* 1.1 *for which $\mathcal{C} < \infty$, where $\mathcal{C}$ is defined by* (1.10)*. Then*

$$1 + \mu\, \{\sigma(\mathcal{C})\}^{(m-1)/2m} \le \omega \le 1 + \nu\, \{\sigma(\mathcal{C})\}^{(m-1)/2m},$$

where μ and ν are positive constants which depend only on m and n, and

$$\sigma(\mathcal{C}) := \begin{cases} \mathcal{C} & \text{for } n < 2, \\ \mathcal{C}\, \{1 + \ln(1+\mathcal{C})\}^{-1} & \text{for } n = 2, \\ \mathcal{C}\, (1+\mathcal{C})^{-(m-1)(n-2)/\{(m-1)n+2\}} & \text{for } n > 2. \end{cases} \tag{1.15}$$

It follows that the size of the blow-up sets Ω and Ω^* may be estimated solely in terms of the critical parameter $\mathcal{C}$. Specifically, if $\mathcal{C} > 0$ then $\omega = \omega^* > 1$, while if $\mathcal{C} = 0$ then $\omega = \omega^* = 1$. This improves results in [15] where, in the case $n = 1$ and under the more restrictive assumptions on the initial and boundary data, estimates of the size of the blow-up set Ω in terms of the parameter defined by (1.8) and kindred parameters were obtained.

Concerning the Cauchy–Neumann problem, our main results on localization and blow-up are as follows.

THEOREM 1.10. *Let u be a weak solution of Problem* 1.2 *and let $\mathcal{C}$ be defined by* (1.11)*. Then $\omega < \infty$ if and only if $\mathcal{C} < \infty$.*

THEOREM 1.11. *Let u be a weak solution of Problem* 1.2 *and $\mathcal{C}$ be specified by* (1.11)*. Define $\ell(r)$ by* (1.9) *and*

$$L(r) := \begin{cases} \mathcal{C} & \text{for } r = 1, \\ \limsup_{t\uparrow T} (T-t)^{1/(m-1)} \left(-r^{n-1} \int_0^t \frac{\partial u^m}{\partial r}(r,s)\, ds\right) & \text{for } r > 1. \end{cases} \tag{1.16}$$

Then, verbatim, conclusions (i) *and* (ii) *of Theorem* 1.7 *hold. Furthermore, if g is nondecreasing, then $u(\cdot, t) \to \infty$ as $t \uparrow T$ uniformly on $(1, r)$ for every $1 < r < \omega$.*

THEOREM 1.12. *Let u_1 and u_2 be two weak solutions of Problem* 1.2 *with corresponding initial data functions $u_{0,i}$, boundary data functions g_i, parameters $\mathcal{C}_i$ defined by* (1.11)*, and functions ℓ_i and L_i defined by* (1.9) *and* (1.16) *for $i = 1, 2$. If*

$$\limsup_{t\uparrow T} (T-t)^{1/(m-1)} \int_0^t \{g_1(s) - g_2(s)\}\, ds \le 0,$$

then $L_1(r) \leq L_2(r)$ *for all* $r \geq 1$. *Moreover, if*

$$(T-t)^{1/(m-1)} \int_0^t \max\{g_1(s) - g_2(s), 0\}\, ds \to 0 \qquad \textit{as } t \uparrow T,$$

then $\ell_1(r) \leq \ell_2(r)$ *for all* $r > 1$.

Theorem 1.13. *Let* u *be a weak solution of Problem* 1.2 *for which* $\mathcal{C} < \infty$, *where* $\mathcal{C}$ *is defined by* (1.11). *Then*

$$1 + \mu \left\{\sigma(\mathcal{C})\right\}^{(m-1)/(m+1)} \leq \omega \leq 1 + \nu \left\{\sigma(\mathcal{C})\right\}^{(m-1)/(m+1)},$$

where μ *and* ν *are positive constants which depend only on* m *and* n, *and*

$$\sigma(\mathcal{C}) := \mathcal{C}\,(1+\mathcal{C})^{-(m-1)(n-1)/\{(m-1)n+2\}}. \tag{1.17}$$

Theorem 1.10 shows that the result of Cortazar and Elgueta [11] for Problem 1.2 in the case $n = 1$ holds for arbitrary $n \geq 1$. Also in the case $n = 1$ it can be obtained under less restrictive assumptions on the boundary and initial data functions. Concurrently, Theorem 1.13 states that the size of the blow-up sets Ω and Ω^* can be estimated solely in terms of the critical parameter $\mathcal{C}$ given by (1.11). As a matter of fact, taking Theorems 1.5, 1.10, and 1.13 together, we see that $\omega = \omega^* = \infty$ when $\mathcal{C} = \infty$, $1 < \omega = \omega^* < \infty$ if $0 < \mathcal{C} < \infty$, and $\omega = \omega^* = 1$ if $\mathcal{C} = 0$. Theorem 1.11 provides the extension of the results in [15] on the behavior of ℓ for the Cauchy–Dirichlet problem in the case $n = 1$ to the Cauchy–Neumann problem and to every $n \geq 1$.

This paper is organized as follows. After this introduction, we present the major tools which we use to prove our theorems. These are special comparison principles, which we establish in section 2, particular self-similar solutions of the porous media equation, whose existence we justify in section 3, and regularity estimates for solutions of (1.1), which we derive in section 4. Since Theorem 1.5 is an easy consequence of one of these estimates, we shall also prove this theorem in section 4. Thereafter, in section 5 we present the bulk of our analysis for Problem 1.1. In a sequence of subsections, we prove, in the following order, Theorem 1.6, Theorem 1.8, Theorem 1.7, Theorem 1.9, and an additional result (Theorem 5.17) on the magnitude of the variables ℓ and L defined by (1.9) and (1.12) for a solution u of the Cauchy–Dirichlet problem. In section 6, we repeat the story for the Cauchy–Neumann problem, treating successively Theorem 1.10, Theorem 1.12, Theorem 1.11 (dependent upon a lemma whose proof we postpone), Theorem 1.13, and a result (Theorem 6.11) on the magnitude of the variables ℓ and L defined by (1.9) and (1.16) for a solution u of Problem 1.2. Finally, in section 7, we comment on the results which can be obtained for the Cauchy–Neumann problem when g is not necessarily nonnegative, and complete the proof of the lemma postponed from section 6.

Using the group-invariance properties of the porous media equation, by a rescaling argument our results also apply to the analogous problems for (1.1) in the domain $(r_0, \infty) \times (0, T)$ for any $r_0 > 0$. One may wonder what is to be expected for the limiting situation with $r_0 = 0$. "Back of an envelope" calculations which we have carried out indicate that the problem of solving (1.1) in $(0, \infty) \times (0, T)$ with the boundary condition $u(0, t) = f(t)$ for $0 < t < T$ and the initial condition $u(r, 0) = u_0(r)$ for $r > 0$ is ill posed when $n > 1$. In particular, if $n > 1$ is an integer then as $r_0 \downarrow 0$, solutions of the problem of (1.1) in $(r_0,\infty)\times(0,T)$ with the boundary condition $u(r_0,t)=f(t)$ for $0<t<T$ and the initial condition $u(r,0)=u_0(r)$ for $r>r_0$ for a function u_0 defined on $(0,\infty)$ converge to the radially symmetric solution of the Cauchy problem for the porous media equation in $\mathbb{R}^n\times(0,T)$ with initial data $u(\boldsymbol{x},0)=u_0(|\boldsymbol{x}|)$. The same destiny appears to await solutions of the problem of (1.1) in $(r_0,\infty)\times(0,T)$ with the boundary condition $-(\partial u^m/\partial r)(r_0,t)=g(t)$ for $0<t<T$ and the initial condition $u(r,0)=u_0(r)$ for $r>r_0$. However, solutions of this problem with the boundary condition $-r_0^{n-1}(\partial u^m/\partial r)(r_0,t)=g(t)$ will converge formally to the solution of the equation $\partial u/\partial t=\Delta u^m+\chi g(t)\delta(\boldsymbol{x})$ in $\mathbb{R}^n\times(0,T)$ with initial data $u(\boldsymbol{x},0)=u_0(|\boldsymbol{x}|)$, where χ is the $(n-1)$-dimensional measure of the surface of the unit ball in $\mathbb{R}^n$.

Our results can also be invoked to study the localization and effective localization of solutions of initial-boundary-value problems for the porous media equation in arbitrary spatial domains in $\mathbb{R}^n$. The only tool needed is an adequate comparison principle. Consider, for instance, the equation $\partial u/\partial t=\Delta u^m$ in $\mathcal{Q}:=D\times(0,T)$ subject to the initial condition $u(\boldsymbol{x},0)=u_0(\boldsymbol{x})$ for $\boldsymbol{x}\in D$ and the boundary condition $u(\boldsymbol{x},t)=f(\boldsymbol{x},t)$ for $(\boldsymbol{x},t)\in\partial D\times(0,T)$, where D is an open subset of $\mathbb{R}^n$, and the complement of $\overline{D}$ is bounded, simply connected, and not empty. By analysis in [12, 13, 29], under the hypotheses that ∂D is smooth, u_0 is a nonnegative function in $L^\infty(D)\cap L^1(D)$, and f is a nonnegative function in $C(\partial D\times[0,T))$, this problem has a unique weak solution, $u\in C(\overline{D}\times(0,\tau])\cap L^\infty(D\times(0,\tau])$ for every $0<\tau<T$, which satisfies a comparison principle. Furthermore, if $\mathbf{0}\notin\overline{D}$ and there exists a finite ζ_0 such that $u_0(\boldsymbol{x})=0$ for all $\boldsymbol{x}\in D$ with $|\boldsymbol{x}|\geq\zeta_0$, then one can define a nondecreasing function $\zeta(t):=\inf\{r:u(\boldsymbol{x},t)=0$ for all $\boldsymbol{x}\in D$ with $|\boldsymbol{x}|\geq r\}$ for $0<t<T$, $\zeta(T)$ by (1.3), and the blow-up set $\Omega:=\{\boldsymbol{x}\in D:\limsup_{t\uparrow T}u(\boldsymbol{x},t)=\infty\}$. Let $M:=\operatorname{ess\,sup}\{u_0(\boldsymbol{x}):\boldsymbol{x}\in D\}$, $r_1:=\min\{|\boldsymbol{x}|:\boldsymbol{x}\in\partial D\}$, and $r_2:=\max\{|\boldsymbol{x}|:\boldsymbol{x}\in\partial D\}$. Define $u_{0,1}(r):=0$ for $r>r_1$, $u_{0,2}(r):=M$ for $r_2<r<\zeta_0$, and $u_{0,2}(r):=0$ for $r\geq\zeta_0$, $f_1(t):=\inf\{f(\boldsymbol{x},s):(\boldsymbol{x},s)\in\partial D\times[t,T)\}$, and $f_2(t):=\max\{\max\{f(\boldsymbol{x},s),M\}:(\boldsymbol{x},s)\in\partial D\times[0,t]\}$ for $0\leq t<T$. In this case, setting $D_i:=\{\boldsymbol{x}\in\mathbb{R}^n:|\boldsymbol{x}|>r_i\}$, there is a radially symmetric solution u_i of the porous media equation in $\mathcal{Q}_i:=D_i\times(0,T)$ satisfying $u_i(\boldsymbol{x},0)=u_{0,i}(|\boldsymbol{x}|)$ for $\boldsymbol{x}\in D_i$, and $u_i(\boldsymbol{x},t)=f_i(t)$ for $(\boldsymbol{x},t)\in\partial D_i\times(0,T)$, for $i=1,2$. Employing the available comparison principle to compare each solution with constant solutions, it can be deduced that $u(\boldsymbol{x},t)\leq f_2(t)$ for all $(\boldsymbol{x},t)\in\mathcal{Q}$ and $u_1(\boldsymbol{x},t)\leq f_1(t)$ for all $(\boldsymbol{x},t)\in\mathcal{Q}_1$; cf. [15]. Subsequently, using the principle to compare the solutions with one another, it can be concluded that $u_1\leq u$ in $\mathcal{Q}$ and $u\leq u_2$ in $\mathcal{Q}_2$. If, by the way, $r_1=r_2$, it suffices to define $f_1(t):=\min\{f(\boldsymbol{x},t):\boldsymbol{x}\in\partial D\}$ and $f_2(t):=\max\{f(\boldsymbol{x},t):\boldsymbol{x}\in\partial D\}$, and the first comparison argument in the foregoing discussion can be skipped. Now, subject to the scaling mentioned above, u_1 and u_2 fall within the scope of our results. Defining $\mathcal{C}_i$ by (1.10) with f replaced by f_i for $i=1,2$ and recalling the notation μ, ν, and σ from Theorem 1.9, we find the following. If $\mathcal{C}_2<\infty$, then $\zeta(T)<\infty$ and $\Omega\subseteq\{\boldsymbol{x}\in D:|\boldsymbol{x}|\leq\varpi_2\}$, where

$$\varpi_2:=r_2\left(1+\nu\left\{\sigma(r_2^{-2m/(m-1)}\mathcal{C}_2)\right\}^{(m-1)/2m}\right).$$

If $\mathcal{C}_1<\infty$, then $\{\boldsymbol{x}\in D:|\boldsymbol{x}|<\varpi_1\}\subseteq\Omega$, where

$$\varpi_1:=r_1\left(1+\mu\left\{\sigma(r_1^{-2m/(m-1)}\mathcal{C}_1)\right\}^{(m-1)/2m}\right);$$

while $\zeta(T)=\infty$ and $\Omega=D$ if $\mathcal{C}_1=\infty$. Similar arguments can be applied to Cauchy–Neumann problems and to problems involving spatial domains which are the exterior of a cylinder and spatial domains which are wedge-shaped. See, for instance, [31] for a discussion of how a solution dependent on a single spatial variable can be deployed to study localization for a problem in a spatial domain given in two dimensions by a quarter-plane.

2. Comparison principles. The principal key to our results is a comparison principle of a kind, first exposed and referred to as a "shifting-comparison principle" by Vázquez [34, 35]. This kind of comparison principle has been further developed in [1, 11, 16]. The comparison principle does not so much concern a solution of the equation but an integral expression involving a solution.

For a solution u of Problem 1.1 this is the variable defined by

$$z(r,t) := \int_r^\infty \kappa(x,r)u(x,t)\,dx \qquad \text{for } (r,t) \in [1,\infty)\times(0,T) \tag{2.1}$$

and

$$z(r,0) := \int_r^\infty \kappa(x,r)u_0(x)\,dx \qquad \text{for } r \in [1,\infty), \tag{2.2}$$

where

$$\kappa(x,r) := \begin{cases} \left(x^{n-1}r^{2-n} - x\right)/(n-2) & \text{if } n \neq 2, \\ x\ln(x/r) & \text{if } n = 2. \end{cases} \tag{2.3}$$

The function κ satisfies the equation

$$x^{n-1}\frac{\partial}{\partial x}(x^{1-n}\kappa) = 1, \tag{2.4}$$

with $\kappa(r,r) = 0$.

Lemma 2.1. *Let z be defined by* (2.1)–(2.3). *Then $z \in C([1,\infty)\times[0,T)) \cap C^{2,1}(Q)$ and*

$$z(1,t) = z(1,0) + \int_0^t f^m(s)\,ds \qquad \textit{for all } 0 < t < T. \tag{2.5}$$

Furthermore,

$$\frac{\partial z}{\partial r} = -r^{1-n}\int_r^\infty x^{n-1}u(x,t)\,dx \le 0, \tag{2.6}$$

$$\frac{\partial}{\partial r}\left(r^{n-1}\frac{\partial z}{\partial r}\right) = r^{n-1}u(r,t) \ge 0, \tag{2.7}$$

and

$$\frac{\partial z}{\partial t} = u^m(r,t) \ge 0 \qquad \textit{for all } (r,t) \in Q. \tag{2.8}$$

Proof. The equality (2.5) follows from the definition of a weak solution of Problem 1.1. The proof of this is identical to the proof of the first lemma in [14] and we omit it. Also let us note that for any $r_0 > 1$ the function

$$\widetilde{u}(r,t) := r_0^{-2/(m-1)}u(r_0 r,t) \tag{2.9}$$

is a weak solution of Problem 1.1 with correspondingly scaled initial data and with boundary data

$$\widetilde{f}(t) := {r_0}^{-2/(m-1)} u(r_0, t). \tag{2.10}$$

Whence, applying (2.5) to the corresponding variable $\widetilde{z}$, reformulating the resulting equality in terms of the original variable, and finally writing r instead of r_0, we deduce

$$z(r,t) = z(r,0) + \int_0^t u^m(r,s)\,ds \quad \text{for all } (r,t) \in Q. \tag{2.11}$$

Clearly then, z is continuous in $[1,\infty) \times [0,T)$. Differentiating (2.1) we find (2.6) and (2.7), while differentiating (2.11) yields (2.8). □

It follows that when n is a positive integer, z is a radially symmetric solution of the equation $\partial z/\partial t = |\Delta z|^{m-1}\Delta z$ in $\{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| > 1\} \times (0,T)$. This equation has been called the *dual porous medium equation* and has been studied in [9, 20]. The relations (2.5) and (2.11) can be interpreted as a principle of conservation of momentum.

The comparison principle involving the variable z is the following.

PROPOSITION 2.2 (comparison principle). *Let u_1 and u_2 be two weak solutions of Problem* 1.1 *with corresponding functions z_i defined by* (2.1)–(2.3) *for $i = 1, 2$. Let*

$$Q_\tau := (1,\infty) \times (0,\tau] \quad \textit{for } 0 < \tau < T. \tag{2.12}$$

Then

$$z_1(r,\tau) - z_2(r,\tau) \le \max\{z_1(x,t) - z_2(x,t) : (x,t) \in \overline{Q_\tau} \setminus Q_\tau\} \tag{2.13}$$

for all $(r,\tau) \in Q$.

Proof. To begin with, suppose that u_1 and u_2 are bounded classical solutions of (1.1) in the closure of Q_τ such that $u_i(r,t) \ge \varepsilon$ for all $(r,t) \in \overline{Q_\tau}$ and $u_i(r,0) = \varepsilon$ for large enough r, for $i = 1, 2$, and some $\varepsilon > 0$. In this case, following the discussion above, it can be shown that

$$w(r,t) := \int_r^\infty \kappa(x,r)(u_1 - u_2)(x,t)\,dx \tag{2.14}$$

is a well-defined bounded $C^{2,1}(\overline{Q_\tau})$ function satisfying the equation

$$\frac{\partial w}{\partial t} = \gamma \frac{\partial^2 w}{\partial r^2} + (n-1)\frac{\gamma}{r}\frac{\partial w}{\partial r}, \tag{2.15}$$

where

$$\gamma(r,t) := m \int_0^1 \{\theta u_1(r,t) + (1-\theta) u_2(r,t)\}^{m-1}\,d\theta \tag{2.16}$$

in Q_τ. Subsequently, applying the classical maximum principle [37] to (2.15),

$$w(r,t) \le \sup\{w(x,s) : (x,s) \in \overline{Q_\tau} \setminus Q_\tau\} \quad \text{for all } (r,t) \in Q_\tau. \tag{2.17}$$

Now, the original weak solutions u_1 and u_2 may be constructed in Q_τ as the limit of a sequence of bounded classical solutions of (1.1) with the previously described properties [18]. Hence, utilizing this construction process, one may obtain (2.17) with w defined by (2.14) for the original solutions. Rewriting (2.17) for this w yields $(z_1 - z_2)(r,t) \leq \max\{(z_1 - z_2)(x,s) : (x,s) \in \overline{Q_\tau} \setminus Q_\tau\}$ for all $(r,t) \in Q_\tau$. This gives the desired result. □

For a weak solution u of Problem 1.2 the role of the variable z defined above is filled by

$$z(r,t) := \int_r^\infty x^{n-1} u(x,t)\,dx \quad \text{for } (r,t) \in [1,\infty) \times (0,T) \tag{2.18}$$

with

$$z(r,0) := \int_r^\infty x^{n-1} u_0(x)\,dx \quad \text{for } r \in [1,\infty). \tag{2.19}$$

This variable has the following properties.

LEMMA 2.3. *Let z be defined by* (2.18) *and* (2.19). *Then $z \in C([1,\infty) \times [0,T)) \cap C^{1,1}(Q)$ and*

$$z(1,t) = z(1,0) + \int_0^t g(s)\,ds \quad \textit{for all } 0 < t < T. \tag{2.20}$$

Furthermore,

$$\frac{\partial z}{\partial r} = -r^{n-1} u(r,t) \leq 0 \tag{2.21}$$

and

$$\frac{\partial z}{\partial t} = -r^{n-1} \frac{\partial u^m}{\partial r} \quad \textit{for all } (r,t) \in Q. \tag{2.22}$$

Proof. The proof of this lemma is similar to the proof of Lemma 2.1. The identity (2.20) may be obtained from the definition of a solution. Moreover, it may be derived under the assumption that

$$g \in L^1(0,\tau) \quad \text{for every } 0 < \tau < T \tag{2.23}$$

and irrespective of the assumption that g is nonnegative. Consequently, since it is known that $\partial u^m/\partial r \in C(Q)$ and $(\partial u^m/\partial r)(r,\cdot) \in L^1(0,\tau)$ for every $(r,\tau) \in Q$, applying a scaling argument similar to that in the proof of Lemma 2.1, we may deduce that

$$z(r,t) = z(r,0) - r^{n-1} \int_0^t \frac{\partial u^m}{\partial r}(r,s)\,ds \quad \text{for all } 0 < t < T. \tag{2.24}$$

The rest of the lemma then follows straightforwardly analogous to Lemma 2.1. □

When $n = 1$ the function z is a solution of $\partial z/\partial t = \nabla \cdot (|\nabla z|^{m-1} \nabla z)$ in $\{x \in \mathbb{R} : |x| > 1\} \times (0,T)$. The identities (2.20) and (2.24) can be interpreted as a principle of conservation of mass.

PROPOSITION 2.4 (comparison principle). *Let u_1 and u_2 be two weak solutions of Problem* 1.2 *with corresponding functions z_i defined by* (2.18) *and* (2.19) *for $i = 1,2$. Then* (2.13) *holds for all $(r,\tau) \in Q$, with Q_τ defined by* (2.12).

Proof. If u_1 and u_2 are classical solutions of (1.1) in the closure of some domain Q_τ and have the other properties stated at the start of the proof of Proposition 2.2, then it can be verified that

$$w(r,t) := \int_r^\infty x^{n-1}(u_1 - u_2)(x,t)\,dx$$

is a classical solution of the equation

$$\frac{\partial w}{\partial t} = \gamma \frac{\partial^2 w}{\partial r^2} + \left\{(1-n)\frac{\gamma}{r} + \frac{\partial \gamma}{\partial r}\right\} \frac{\partial w}{\partial r},$$

where γ is once more defined by (2.16) in Q_τ. The proof of this theorem may be subsequently completed similarly to that of the previous one. □

3. Self-similar solutions.

3. Self-similar solutions. The second major tool which we use in our analysis is the existence of self-similar solutions of (1.1).

The first class of self-similar solutions of (1.1) which we consider is the generalization of the waiting-time solution of the one-dimensional porous media equation discovered by Kalashnikov [21].

PROPOSITION 3.1. *There exists a strictly decreasing, continuous function ρ defined on $(0,1]$ with $\rho(1) = 0$, such that for any $a > 0$ and $\tau \geq T$ the function*

$$U(r,t;a,\tau) := \begin{cases} (\tau - t)^{-1/(m-1)} a^{2/(m-1)} \rho(r/a) & \textit{for } r < a, \\ 0 & \textit{otherwise} \end{cases} \tag{3.1}$$

is a weak solution of Problems 1.1 *and* 1.2 *with appropriate initial and boundary data.*

Proof. Making the formal substitution of (3.1) into (1.1), one finds that ρ satisfies the ordinary differential equation

$$(m-1)x^{1-n}(x^{n-1}(\rho^m)')' = \rho. \tag{3.2}$$

Correspondingly, the existence of the self-similar solution (3.1) can be proven by manipulation in the definition of a weak solution of Problems 1.1 and 1.2 once the next lemma is established. □

LEMMA 3.2. *Equation* (3.2) *has a unique positive solution ρ for $x < 1$ satisfying*

$$\rho(1) = (\rho^m)'(1) = 0 \tag{3.3}$$

which is extendible onto $(0,1)$ and strictly decreasing in this interval. Furthermore,

$$(\rho^m)'(x) \sim -\lambda_0 x^{1-n} \qquad \textit{as } x \downarrow 0 \tag{3.4}$$

and

$$\rho^{(m-1)/2}(x) \sim \lambda_1(1-x) \qquad \textit{as } x \uparrow 1, \tag{3.5}$$

where λ_0 is a positive number and $\lambda_1 := \{(m-1)/2m(m+1)\}^{1/2}$.

Proof. To prove the existence, we shall use a contraction mapping argument which is commonly applied for treating this kind of initial-value problem [5, 6, 17, 24, 28]. Multiplying (3.2) by $2x^{2(n-1)}(\rho^m)'(x)$, integrating from x to 1, and using (3.3) gives

$$\{x^{n-1}(\rho^m)'(x)\}^2 = \frac{2m}{(m-1)(m+1)} x^{2(n-1)} \rho^{m+1}(x) + \frac{4m(n-1)}{(m-1)(m+1)} \int_x^1 y^{2n-3} \rho^{m+1}(y)\,dy.$$

Hence a solution ρ of (3.2), (3.3) satisfies $(\rho^m)'(x) < 0$ and

$$(\rho^{(m-1)/2})'(x) = -\lambda_1 \left\{ 1 + 2(n-1)x^{2(1-n)} \int_x^1 y^{2n-3} \left(\frac{\rho(y)}{\rho(x)} \right)^{m+1} dy \right\}^{1/2}$$

for all $x < 1$. Subsequently, integrating from x to 1 a second time,

$$\begin{aligned} &\rho^{(m-1)/2}(x) \\ &\quad = \lambda_1 \int_x^1 \left\{ 1 + 2(n-1)y^{2(1-n)} \int_y^1 \eta^{2n-3} \left(\frac{\rho(\eta)}{\rho(y)} \right)^{m+1} d\eta \right\}^{1/2} dy. \end{aligned} \tag{3.6}$$

It follows that

$$\begin{aligned} \lambda_1(1-x) &\leq \rho^{(m-1)/2}(x) \\ &\leq \lambda_1 \int_x^1 \left\{ 1 + 2(n-1)y^{2(1-n)} \int_y^1 \eta^{2n-3}\, d\eta \right\}^{1/2} dy \\ &= \lambda_1 \int_x^1 y^{1-n}\, dy \leq \lambda_1(1-x)x^{1-n} \end{aligned} \tag{3.7}$$

for all $x < 1$. Now let $\mathcal{S}$ denote the set of functions $\phi \in C([1-\delta, 1])$ such that $\lambda_1 \leq \phi(x) \leq \lambda_1(1-\delta)^{1-n}$ for all $1-\delta \leq x \leq 1$ for some $0 < \delta < 1$, and define the mapping M on $\mathcal{S}$ by

$$\begin{aligned} &M(\phi)(x) \\ &:= \frac{\lambda_1}{1-x} \int_x^1 \left\{ 1 + 2(n-1)y^{2(1-n)} \int_y^1 \eta^{2n-3} \left(\frac{(1-\eta)\phi(\eta)}{(1-y)\phi(y)} \right)^{2(m+1)/(m-1)} d\eta \right\}^{1/2} dy \end{aligned}$$

for all $1-\delta \leq x < 1$. It is possible to show that if δ is chosen small enough, M maps $\mathcal{S}$ into itself and moreover is a contraction with respect to the $C([1-\delta, 1])$-norm. Therefore, since $\mathcal{S}$ is closed with respect to this norm, M has a unique fixed point $\phi \in \mathcal{S}$ by the contraction mapping principle. Setting $\rho(x) = (1-x)^{2/(m-1)}\phi^{2/(m-1)}(x)$, it follows that there is a unique function ρ satisfying (3.6) for $1-\delta < x < 1$ for small enough δ. Whence, retracing the above argument, problem (3.2), (3.3) has a unique positive solution ρ for $1-\delta < x < 1$ for small enough δ. In view of the a priori estimate (3.7), this solution is continuously and uniquely extendible onto $(0, 1)$, by standard theory for ordinary differential equations.

To prove (3.4) we multiply (3.2) by $x^{n-1}/(m-1)$ and integrate from x to 1, to obtain

$$-(\rho^m)'(x) = \frac{1}{m-1} x^{1-n} \int_x^1 y^{n-1}\rho(y)\, dy. \tag{3.8}$$

Consequently, if we can show that

$$\int_0^1 y^{n-1}\rho(y)\, dy < \infty, \tag{3.9}$$

then (3.4) is proven. To verify (3.9) in the case $n \leq 2$, we use the monotonicity of ρ. Substituting $\rho(y) < \rho(x)$ in (3.8) yields

$$-(\rho^{m-1})'(x) < \frac{1}{m} x^{1-n} \int_x^1 y^{n-1}\, dy < x^{-1}$$

for any $0 < x < 1$. Hence, integrating from y to 1, we deduce that $\rho(y) < |\ln y|^{1/(m-1)}$ for every $0 < y < 1$. This is sufficient to confirm (3.9) when $n \leq 2$. To verify (3.9) in the case $n > 2$ we have to work a little harder. Fix $0 < x < 1$, and define $A := \sup\{y^{(n-2)/m}\rho(y) : x < y < 1\}$. Substituting $\rho(y) \leq Ay^{(2-n)/m}$ in the right-hand side of (3.8) and integrating give

$$\rho^m(y) < \frac{m}{(m-1)(n-2)\{(m-1)n+2\}} y^{2-n} A$$

for any $x < y < 1$. Hence, multiplying this inequality by y^{n-2}/A and subsequently optimizing the left-hand side of the resulting expression, we obtain $A^{m-1} \leq m/(m-1)(n-2)\{(m-1)n+2\}$. We therefore conclude that

$$\rho(y) \leq \left(\frac{m}{(m-1)(n-2)\{(m-1)n+2\}}\right)^{1/(m-1)} y^{(2-n)/m}$$

for all $x < y < 1$, irrespective of the value of $x > 0$. This provides (3.9) when $n > 2$. The relation (3.5) follows from (3.7). □

In the case $n = 1$, the expression (3.6) gives the function ρ explicitly as

$$\rho(x) = \left\{\frac{m-1}{2m(m+1)}(1-x)^2\right\}^{1/(m-1)}. \tag{3.10}$$

This gives rise to the waiting-time solution of the porous media equation documented in [21].

The second self-similar solution of (1.1) which we consider is the instantaneous point-source solution. This solution was discovered in the cases $n = 1$ and $n = 3$ by Zel'dovich and Kompaneets [38] and in the general case by Barenblatt [7] and Pattle [25]. It is often referred to as the Barenblatt–Pattle or Barenblatt solution [3, 4, 27, 36].

Proposition 3.3. *For any $a > 0$ and τ the function*

$$U(r,t;a,\tau) := \begin{cases} (t-\tau)^{-1/(m-1)}\xi(t)^{2/(m-1)}\rho(r/\xi(t)) & \text{for } t > \tau \text{ and } r < \xi(t), \\ 0 & \text{otherwise}, \end{cases} \tag{3.11}$$

where

$$\xi(t) := a(t-\tau)^{1/\{(m-1)n+2\}} \tag{3.12}$$

and

$$\rho(x) := \left[\frac{m-1}{2m\{(m-1)n+2\}}\left(1-x^2\right)\right]^{1/(m-1)}, \tag{3.13}$$

is a weak solution of Problems 1.1 *and* 1.2 *with appropriate initial and boundary data.*

In the case that n is a positive integer, U is formally a radially symmetric solution of

$$\frac{\partial u}{\partial t} = \Delta u^m + \chi a^{\{(m-1)n+2\}/(m-1)}\delta(\boldsymbol{x})\delta(t-\tau),$$

where

$$\chi := \int_{\{\boldsymbol{x}\in\mathbb{R}^n:|\boldsymbol{x}|<1\}} \rho(|\boldsymbol{y}|)\, d\boldsymbol{y},$$

in $\mathbb{R}^n \times \mathbb{R}$. The function U is correspondingly singular for $(r,t) = (0,\tau)$. However, since this point lies outside our study domain Q, this is of no consequence for our analysis.

4. Regularity estimates. Our final principal tool comprises two estimates of the regularity of solutions of Problems 1.1 and 1.2. The first of these is an elaboration of a well-known regularity result for the porous media equation in one spatial dimension due to Aronson [2]. Corresponding results for more general quasilinear parabolic equations of second order have been derived in [8, 16, 22].

PROPOSITION 4.1. *Given any* $1 \le r_0 < r_1$ *there exists a constant* K *which depends only on* m, n, *and* $\delta := (r_1 - r_0)/r_0$ *such that for any weak solution* u *of Problem* 1.1 *or* 1.2 *there holds*

$$\left|u^{m-1}(x,\tau) - u^{m-1}(y,\tau)\right| \le K\left(\tau^{-1}M^{m-1} + r_0^{-2}M^{2(m-1)}\right)^{1/2}|x-y|$$

for all $x \ge r_1$, $y \ge r_1$ *and* $0 < \tau < T$, *where* $M := \sup\{u(r,t) : (r,t) \in (r_0,\infty) \times (0,\tau]\}$.

This result may be obtained by following the method used in [2]. Full documentation is supplied in [19]. The only embellishment on the result appearing in [19] is the stated dependence on r_0 and M. This may be obtained from the invariance of (1.1) under scaling; see, for instance, the proof of the next result.

PROPOSITION 4.2. *Given any* $1 \le r_0 < r_1$ *there exists a constant* K *which depends only on* m, n, *and* $\delta := (r_1 - r_0)/r_0$ *such that for any weak solution* u *of Problem* 1.1 *or* 1.2 *there holds*

$$u(r,\tau) \le Kr^{-q}\left(\tau^{-1}N^2 + r_0^{-\{(m-1)n+2\}}N^{m+1}\right)^{1/(m+1)}$$

for all $r \ge r_1$ *and* $0 < \tau < T$, *where*

$$q := 2(n-1)/(m+1)$$

and $N := \max\{z(r_0,t) : 0 \le t \le \tau\}$ *with* z *defined by* (2.18).

Proof. We adapt the Bernstein argument as applied in [2, 16]. Fix $1 \le r_0 < r_1 < r_2$ and $0 < \tau < T$. Recall the definition of Q_τ by (2.12). Since any weak solution of Problem 1.1 or 1.2 may be constructed in Q_τ as the limit of a sequence of positive classical $C(\overline{Q_\tau})$ solutions of (1.1), without loss of generality we may suppose that u is as such. Moreover, because (1.1) is invariant under the transformation $u(r,t) \mapsto au(br, a^{m-1}b^2t)$ for any positive numbers a and b, without loss of generality we may suppose that $r_0 = 1$ and $N = 1$. Note that by (2.21) the latter means that $z(r,t) \le 1$ for all $(r,t) \in Q_\tau$. Consider now the function

$$p(r,t) := u(r,t)/\psi(z(r,t)), \tag{4.1}$$

where $\psi \in C^2([0,1])$ is such that $\psi(\eta) > 0$ and $(\psi^m)''(\eta) < 0$ for all $0 \le \eta \le 1$; for instance, $\psi(\eta) := \{(1+\eta)(2-\eta)\}^{1/m}$. Differentiating (4.1) with respect to t and successively using (1.1) to eliminate $\partial u/\partial t$, (2.22) to eliminate $\partial z/\partial t$, (4.1) to eliminate u, and (2.21) to eliminate $\partial z/\partial r$, it can be verified that p satisfies

$$\begin{aligned}\frac{\partial p}{\partial t} &= \psi^{m-1}\frac{\partial^2 p^m}{\partial r^2} + \{(n-1)r^{-1}\psi^{m-1} - 2r^{n-1}(\psi^m)'p\}\frac{\partial p^m}{\partial r} \\ &\quad - 2(n-1)r^{n-2}(\psi^m)'p^{m+1} + r^{2n-2}\psi(\psi^m)''p^{m+2}\end{aligned} \tag{4.2}$$

in Q_τ. Set $r_3 := r_2 + \delta$ and consider next the function

$$w(r,t) := t^{1/(m+1)} r^q \xi(r) p(r,t), \tag{4.3}$$

where $\xi \in C^2([r_0, r_3])$ is such that $\xi(r_0) = \xi(r_3) = 0$, $0 < \xi(r) \leq 1$ for $r_0 < r < r_3$, and $\xi(r) = 1$ for $r_1 \leq r \leq r_2$. By this construction, w is positive in $(r_0, r_3) \times (0, \tau]$ and vanishes on $[r_0, r_3] \times [0, \tau] \setminus (r_0, r_3) \times (0, \tau]$. Hence, a maximum of w in $[r_0, r_3] \times [0, \tau]$ must lie in $(r_0, r_3) \times (0, \tau]$, and at such a point

$$\frac{\partial w^m}{\partial r} = 0, \quad \frac{\partial w}{\partial t} \geq 0, \quad \text{and} \quad \frac{\partial^2 w^m}{\partial r^2} \leq 0. \tag{4.4}$$

Setting (4.3) in (4.4), this implies

$$\frac{\partial p^m}{\partial r} = -m\xi^{-1} p^m \{q r^{-1} \xi + \xi'\}, \tag{4.5}$$

$$\frac{\partial p}{\partial t} \geq -\frac{1}{m+1} t^{-1} p, \tag{4.6}$$

and

$$\frac{\partial^2 p^m}{\partial r^2} \leq m\xi^{-2}\{q(mq+1)r^{-2}\xi^2 + 2mqr^{-1}\xi\xi' + (m+1)(\xi')^2 - \xi\xi''\}p^m. \tag{4.7}$$

Thus, substituting (4.5)–(4.7) in (4.2) and thereafter using (4.3) to eliminate p, at a maximum of w in $[r_0, r_3] \times [0, \tau]$ there holds

$$-\psi(\psi^m)'' w^{m+1} \leq (\psi^m)' \Xi_1 t^{1/(m+1)} w^m + \psi^{m-1} \Xi_2 t^{2/(m+1)} w^{m-1} + \Xi_3,$$

where

$$\Xi_1 := 2r^{-(m-1)q/2}\{(mq - n + 1)r^{-1}\xi + m\xi'\},$$

$$\Xi_2 := mr^{-(m-1)q}\{q(mq - n + 2)r^{-2}\xi^2 + (2mq - n + 1)r^{-1}\xi\xi' + (m+1)(\xi')^2 - \xi\xi''\},$$

and

$$\Xi_3 := \xi^{m+1}/(m+1).$$

It follows that there are positive constants K_1, K_2, and K_3 which depend only on m, n, and $\delta = r_1 - r_0 = r_3 - r_2$ such that

$$W^{m+1} \leq K_1 \tau^{1/(m+1)} W^m + K_2 \tau^{2/(m+1)} W^{m-1} + K_3, \tag{4.8}$$

where W denotes the maximal value of w in $[r_0, r_3] \times [0, \tau]$. Now, by Young's inequality, there is a constant K_4 which depends only on m and K_1 such that

$$K_1 \tau^{1/(m+1)} W^m \leq W^{m+1}/4 + K_4 \tau, \tag{4.9}$$

and there is a constant K_5 which depends only on m and K_2 such that

$$K_2 \tau^{2/(m+1)} W^{m-1} \leq W^{m+1}/4 + K_5 \tau. \tag{4.10}$$

Plugging (4.9) and (4.10) into (4.8) yields $W \leq K_6(1+\tau)^{1/(m+1)}$ with $K_6^{m+1} := 2\max\{K_3, K_4+K_5\}$. This means that $w(r,t) \leq K_6(1+\tau)^{1/(m+1)}$ for all $(r,t) \in [r_0, r_3] \times [0,\tau]$. Hence, in particular, by (4.3) and the properties of ξ, $p(r,\tau) \leq K_6 r^{-q}(\tau^{-1}+1)^{1/(m+1)}$ for all $r_1 \leq r \leq r_2$. Using (4.1), this gives

$$u(r,\tau) \leq K r^{-q}(\tau^{-1}+1)^{1/(m+1)} \quad \text{for all } r_1 \leq r \leq r_2$$

with $K := K_6 \max\{\psi(\eta) : 0 \leq \eta \leq 1\}$. In view of the arbitrariness of r_2, this provides the required result. □

Note that it follows from Proposition 4.2 that for a radially symmetric solution u of the porous media equation in $\mathbb{R}^n \times (0,T)$ we have the estimate

$$\|u(\cdot,t)\|_{L^\infty(E_{2R})} \leq K\left(R^{-n}\|u\|_{L^\infty(0,t;L^1(E_R))} + R^{-q}t^{-1/(m+1)}\|u\|^{2/(m+1)}_{L^\infty(0,t;L^1(E_R))}\right),$$

where $E_R := \{\boldsymbol{x} \in \mathbb{R}^n : |\boldsymbol{x}| > R\}$, for any $R > 0$ and $0 < t < T$, with a constant K which depends only on m and n.

As well as being of some interest in its own right, Proposition 4.2 provides a means to prove Theorem 1.5.

Proof of Theorem 1.5. Suppose that $r \in \Omega$. Then, by definition, $\sup\{u(r,t) : 0 < t < T\} = \infty$. Let $1 < r_0 < r$. Proposition 4.2 subsequently implies that $\sup\{z(r_0,t) : 0 \leq t < T\} = \infty$, where z is defined by (2.18). Hence, $r_0 \in \Omega^*$ by definition. This means that $(1,r) \subseteq \Omega^*$ for all $r \in \Omega$. Combining this deduction with (1.6) and (1.7) gives the theorem. □

5. The Cauchy–Dirichlet problem.

5.1. Notation. Let u denote a given weak solution of Problem 1.1 with initial data function u_0 and boundary data function f. Note the next result which can be proven following analysis in [15] or alternatively, by adapting the proof of the analogous result for the Cauchy–Neumann problem presented in section 6.

LEMMA 5.1. *Suppose that f is nondecreasing and $u_0 \equiv 0$. Then $u(r,t) \geq u(r_0,t_0)$ for all $1 < r \leq r_0$ and $0 < t_0 \leq t < T$.*

Recall the definitions (1.4)–(1.6) and (1.9) and, for the Cauchy–Dirichlet problem, the definition of $\mathcal{C}$ by (1.10), L by (1.12), and z by (2.1)–(2.3). For completeness, define

$$z(r,T) := \lim_{t\uparrow T} z(r,t) \tag{5.1}$$

for $r \geq 1$, which limit exists by (2.5) and (2.11). By (2.5) and (2.11) the definition of L is equivalent to

$$L(r) = \limsup_{t\uparrow T}\,(T-t)^{1/(m-1)}\,z(r,t) \quad \text{for } r \geq 1. \tag{5.2}$$

Let

$$v(r,t) := (T-t)^{1/(m-1)}u(r,t) \quad \text{for } (r,t) \in Q, \tag{5.3}$$

and note that

$$\ell(r) = \limsup_{t\uparrow T} v(r,t) \quad \text{for } r > 1.$$

For convenience we shall denote the variable z defined for an arbitrary solution of the Cauchy–Dirichlet problem for (1.1) by (2.1)–(2.3) and (5.1) with a capital letter when it applies to a self-similar solution. Thus we set

$$Z(r,t;a,\tau) := \int_r^\infty \kappa(x,r)U(x,t;a,\tau)\,dx \tag{5.4}$$

for any $(r,t) \in [1,\infty) \times [0,T)$. Furthermore, we shall introduce a subscript w to indicate when this concerns the self-similar solution of waiting-time type. Substituting (2.3) and (3.1) in (5.4), using (3.2), and thereafter applying partial integration to simplify the resulting expression reveal that

$$Z_w(r,t;a,\tau) = (m-1)a^{2m/(m-1)}(\tau-t)^{-1/(m-1)}\rho^m(r/a)$$

for all $t < \tau$ and $r < a$. Let

$$\mathcal{A}(c) := (m-1)c^{2m/(m-1)}\rho^m(1/c). \tag{5.5}$$

Since ρ has the properties stated in Lemma 3.2, this defines an unbounded, strictly increasing, continuously differentiable function $\mathcal{A}$ on $[1,\infty)$ with $\mathcal{A}(1) = 0$. Subsequently we can state that

$$Z_w(r,t;a,\tau) = \begin{cases} (\tau-t)^{-1/(m-1)}r^{2m/(m-1)}\mathcal{A}(a/r) & \text{if } r < a, \\ 0 & \text{otherwise} \end{cases}$$

for any $(r,t) \in [1,\infty) \times [0,T)$.

Analogous to what we have done for the waiting-time solution, we define the variable $Z_s(r,t;a,\tau)$ for the instantaneous point-source solution. Substituting (2.3), (3.11), and (3.13) in (5.4), noting that $\rho = -\{(m-1)n+2\}x^{-1}(\rho^m)'$, and thereafter applying partial integration yield

$$\begin{aligned} &Z_s(r,t;a,\tau) \\ &\quad = \{(m-1)n+2\}(t-\tau)^{-1/(m-1)}\xi(t)^{\{(m-1)n+2\}/(m-1)}r^{2-n}\int_{r/\xi(t)}^1 x^{n-3}\rho^m(x)\,dx \end{aligned}$$

for all $t > \tau$ and $r < \xi(t)$, where $\xi(t)$ is given by (3.12). Hence, if we define

$$\mathcal{B}(c) := \{(m-1)n+2\}c^{\{(m-1)n+2\}/(m-1)}\int_{1/c}^1 x^{n-3}\rho^m(x)\,dx, \tag{5.6}$$

we can state that

$$Z_s(r,t;a,\tau) = \begin{cases} (t-\tau)^{-1/(m-1)}r^{2m/(m-1)}\mathcal{B}(\xi(t)/r) & \text{if } t > \tau \text{ and } r < \xi(t), \\ 0 & \text{otherwise} \end{cases}$$

for all $(r,t) \in \overline{Q}$, where $\mathcal{B}$ is defined, continuously differentiable, strictly increasing, and unbounded on $[1,\infty)$ and such that $\mathcal{B}(1) = 0$.

Let β and α denote the inverse of the functions $\mathcal{A}$ and $\mathcal{B}$, respectively. Thus α and β are strictly increasing, continuous, and unbounded functions on $[0,\infty)$ such that $\alpha(0) = \beta(0) = 1$. Finally, define the strictly increasing function $\gamma \in C([0,\infty))$ with $\gamma(0) = 1$ by

$$\{\gamma(\mathcal{C})\}^{2m/(m-1)}\,\mathcal{A}(\alpha(\mathcal{C})/\gamma(\mathcal{C})) = 2^{-1/(m-1)}\mathcal{C}. \tag{5.7}$$

These functions have the following property.

LEMMA 5.2. *There are positive constants μ and ν such that* $\max\{\beta(\mathcal{C}),\gamma(\mathcal{C})\} \geq 1+\mu\{\sigma(\mathcal{C})\}^{(m-1)/2m}$ *and* $\alpha(\mathcal{C}) \leq 1+\nu\{\sigma(\mathcal{C})\}^{(m-1)/2m}$ *for all* $\mathcal{C} \geq 0$, *where* σ *is defined by* (1.15).

Proof. In view of the known properties of α, β, and γ, to prove the lemma it is enough to show that

$$\alpha(\mathcal{C})-1 \sim \alpha_0\,\{\sigma(\mathcal{C})\}^{(m-1)/2m} \qquad \text{as } \mathcal{C}\downarrow 0, \tag{5.8}$$

$$\alpha(\mathcal{C})-1 \sim \alpha_1\,\{\sigma(\mathcal{C})\}^{(m-1)/2m} \qquad \text{as } \mathcal{C}\to\infty, \tag{5.9}$$

$$\beta(\mathcal{C})-1 \sim \beta_1\,\{\sigma(\mathcal{C})\}^{(m-1)/2m} \qquad \text{as } \mathcal{C}\to\infty, \tag{5.10}$$

and

$$\gamma(\mathcal{C})-1 \sim \gamma_0\,\{\sigma(\mathcal{C})\}^{(m-1)/2m} \qquad \text{as } \mathcal{C}\downarrow 0 \tag{5.11}$$

for some positive numbers α_0, α_1, β_1, and γ_0. To verify (5.8), note that combining (5.5) and (3.5) there holds

$$\mathcal{A}(c) \sim \mathcal{A}_0(c-1)^{2m/(m-1)} \qquad \text{as } c\downarrow 1 \tag{5.12}$$

for some constant $\mathcal{A}_0>0$. This readily yields (5.8) with $\alpha_0=\mathcal{A}_0^{(1-m)/2m}$. Next, with regard to (5.9), we see that combination of (5.5) and (3.4) implies

$$\mathcal{A}(c) \sim \mathcal{A}_1 c^{2m/(m-1)}\int_{1/c}^{1} x^{1-n}\,dx \qquad \text{as } c\to\infty$$

for some constant $\mathcal{A}_1>0$, whence, by a change of variables,

$$\mathcal{C} \sim \mathcal{A}_1\,\{\alpha(\mathcal{C})\}^{2m/(m-1)}\int_{1}^{\alpha(\mathcal{C})} y^{n-3}\,dy \qquad \text{as } \mathcal{C}\to\infty. \tag{5.13}$$

After suitable manipulation this shows that $\alpha(\mathcal{C}) \sim \alpha_1\mathcal{C}^{(m-1)/2m}$ with $\alpha_1=\{(2-n)/\mathcal{A}_1\}^{(m-1)/2m}$ when $n<2$, $\alpha(\mathcal{C}) \sim \alpha_1(\mathcal{C}/\ln\mathcal{C})^{(m-1)/2m}$ with $\alpha_1=\{2m/(m-1)\mathcal{A}_1\}^{(m-1)/2m}$ when $n=2$, and $\alpha(\mathcal{C}) \sim \alpha_1\mathcal{C}^{(m-1)/\{(m-1)n+2\}}$ with $\alpha_1=\{(n-2)/\mathcal{A}_1\}^{(m-1)/\{(m-1)n+2\}}$ when $n>2$. Together this is equivalent to (5.9). To verify (5.10), observe that substituting (3.13) in (5.6) there holds

$$\mathcal{B}(c) \sim \mathcal{B}_1 c^{\{(m-1)n+2\}/(m-1)}\int_{1/c}^{1} x^{n-3}\,dx \qquad \text{as } c\to\infty$$

for some constant $\mathcal{B}_1>0$, whence, changing variables,

$$\mathcal{C} \sim \mathcal{B}_1\,\{\beta(\mathcal{C})\}^{\{(m-1)n+2\}/(m-1)}\int_{1}^{\beta(\mathcal{C})} y^{1-n}\,dy \qquad \text{as } \mathcal{C}\to\infty.$$

The relation (5.10) subsequently follows by a similar argument to that used to deduce (5.9) from (5.13). One finds $\beta_1=\{(2-n)/\mathcal{B}_1\}^{(m-1)/2m}$ when $n<2$, $\beta_1=\{2m/(m-1)\mathcal{B}_1\}^{(m-1)/2m}$ when $n=2$, and $\beta_1=\{(n-2)/\mathcal{B}_1\}^{(m-1)/\{(m-1)n+2\}}$ when $n>2$. Finally, to obtain (5.11), we observe that by (5.7) and (5.12), $\mathcal{A}_0\{\alpha(\mathcal{C})-\gamma(\mathcal{C})\}^{2m/(m-1)} \sim 2^{-1/(m-1)}\mathcal{C}$ as $\mathcal{C}\downarrow 0$. Together with (5.8) this gives (5.11) with $\gamma_0=(1-2^{-1/2m})\alpha_0$. □

5.2. The proof of Theorem 1.6. In this subsection we essentially prove that localization, effective localization, and metastable localization occur if and only if $\mathcal{C} < \infty$. We use the next five lemmas.

LEMMA 5.3. *Let u_1 and u_2 be any two weak solutions of Problem* 1.1 *with corresponding functions $u_{0,i}$, f_i, L_i, and z_i for $i = 1, 2$. Suppose that* (1.13) *holds. Then $L_1(r) \leq L_2(r)$ for all $r \geq 1$. Moreover, if* $\sup\{z_1(1,t) - z_2(1,t) : 0 \leq t < T\} < \infty$, *then $z_1(r,T) < \infty$ for all $r \geq 1$ such that $z_2(r,T) < \infty$.*

Proof. By Proposition 2.2 and Lemma 2.1,

$$z_1(r,t) \leq z_2(r,t) + z_1(1,\tau) + \max\{(z_1 - z_2)(1,s) : \tau \leq s \leq t\} \tag{5.14}$$

for all $r \geq 1$ and $0 < \tau < t < T$. As a consequence,

$$\begin{aligned}(T-t)^{1/(m-1)} z_1(r,t) &\leq (T-t)^{1/(m-1)} z_2(r,t) + (T-t)^{1/(m-1)} z_1(1,\tau) \\ &\quad + \max\{\max\{(T-s)^{1/(m-1)}(z_1 - z_2)(1,s) : \tau \leq s \leq t\}, 0\}\end{aligned} \tag{5.15}$$

for all $r \geq 1$ and $0 < \tau < t < T$. Letting $t \uparrow T$ and thereafter $\tau \uparrow T$ in (5.15) yields the main conclusion of the lemma. The subsidiary conclusion may be obtained by passing to the limit $t \uparrow T$ in (5.14). □

LEMMA 5.4. *If $c > 1$ is such that $\mathcal{A}(c) > \mathcal{C}$, then*

$$L(r) \leq r^{2m/(m-1)} \mathcal{A}(\max\{c/r, 1\}) \quad \textit{for all } r \geq 1$$

and $z(r,T) < \infty$ for all $r \geq c$.

Proof. Taking $z_1 = z$ and $z_2 = Z_w(\cdot, \cdot; c, T)$, this lemma is a corollary of the previous one. □

LEMMA 5.5. *If $c > 1$ is such that $\mathcal{B}(c) < \mathcal{C}$, then $z(r,T) = \infty$ for all $1 < r < c$.*

Proof. By definition, there exists a sequence of values $\{t_i\}_{i=1}^{\infty} \subset (0,T)$ such that $t_i \to T$ as $i \to \infty$, and $(T - t_i)^{1/(m-1)} z(1,t_i) > \mathcal{B}(c)$ for any $i \geq 1$. Fix i and set $a_i := c(T - t_i)^{-1/\{(m-1)n+2\}}$. Then $z(1,t) \geq z(1,t_i) > (T - t_i)^{-1/(m-1)} \mathcal{B}(c) = Z_s(1,T; a_i, t_i) \geq Z_s(1,t; a_i, t_i)$ for all $t_i < t < T$, while $z(r,t) \geq 0 = Z_s(r,t; a_i, t_i)$ for all $r \geq 1$ and $0 \leq t \leq t_i$. Proposition 2.2 subsequently states that $z(r,t) \geq Z_s(r,t; a_i, t_i)$ for all $(r,t) \in Q$. This gives

$$z(r,T) \geq Z_s(r,T; a_i, t_i) = (T - t_i)^{-1/(m-1)} r^{2m/(m-1)} \mathcal{B}(c/r)$$

for all $1 < r < c$. Letting $i \to \infty$ yields the lemma. □

LEMMA 5.6. *Suppose that $L(r_0) < \infty$ for some $r_0 \geq 1$. Then for any $r > r_0$,*

$$\sup\left\{\int_r^\infty x^{n-1} v(x,t)\,dx : 0 < t < T\right\} < \infty.$$

Moreover, if $z(r_0,T) < \infty$, then $r \notin \Omega^*$.

Proof. By (2.4),

$$\begin{aligned}z(r_0,t) &= \int_{r_0}^r \kappa(x,r_0) u(x,t)\,dx + \int_r^\infty \{x^{1-n}\kappa(x,r_0)\} x^{n-1} u(x,t)\,dx \\ &\geq r^{1-n}\kappa(r,r_0) \int_r^\infty x^{n-1} u(x,t)\,dx\end{aligned} \tag{5.16}$$

for all $0 < t < T$. Multiplying (5.16) by $(T-t)^{1/(m-1)}$ and letting $t \uparrow T$ yields the first conclusion of the lemma, while letting $t \uparrow T$ in (5.16) directly gives the second conclusion. □

LEMMA 5.7. *If $z(r,T) = \infty$ for some $r > 1$, then $r \in \Omega$.*

Proof. This follows immediately from (2.11). □

With the last four lemmas it is easy to prove Theorem 1.6. Suppose that $\mathcal{C} < \infty$. Then by Lemma 5.4, $z(r,T) < \infty$ for all $r > \alpha(\mathcal{C})$. This implies that $\omega^* < \infty$ by Lemma 5.6 and (1.6). On the other hand, if $\mathcal{C} = \infty$, then by Lemma 5.5, $z(r,T) = \infty$ for all $r > 1$. Thus, by Lemma 5.7 and (1.6), $\omega = \infty$. Recalling Theorem 1.5 gives the required result.

5.3. The proof of Theorem 1.8. To establish our fourth theorem, we employ four more lemmas.

LEMMA 5.8. *Suppose that $L(r_0) < \infty$ for some $r_0 \geq 1$. Then*

$$L(r) \leq r^{2m/(m-1)}\mathcal{A}(\max\{r_1/r, 1\}) \quad \text{for all } r \geq r_0, \tag{5.17}$$

and $z(r,T) < \infty$ for all $r > r_1$, where

$$r_1 := r_0\alpha(r_0^{-2m/(m-1)}L(r_0)). \tag{5.18}$$

Proof. We apply the scaling technique employed in the proof of Lemma 2.1. The function $\widetilde{u}$ given by (2.9) is a weak solution of Problem 1.1 with correspondingly scaled initial values, with boundary values (2.10), and with the parameter defined by (1.10) taking the value $r_0^{-2m/(m-1)}L(r_0)$. The present lemma may subsequently be obtained by applying Lemma 5.4 to $\widetilde{u}$ and interpreting the conclusions in terms of the original solution. □

LEMMA 5.9. *If $\mathcal{C} = \infty$, then $L(r) = \ell(r) = \infty$ for all $r > 1$.*

Proof. If $\mathcal{C} = \infty$, then $z(r,T) = \infty$ for all $r > 1$ by Lemma 5.5. Subsequently, by Lemma 5.8, $L(r) = \infty$ for all $r > 1$. However, by the definition of ℓ and L,

$$L(r) \leq (m-1)\ell^m(r) \quad \text{for every } r > 1. \qquad \square \tag{5.19}$$

LEMMA 5.10. *If $\mathcal{C} < \infty$ then $v^{m-1} \in L^\infty(\tau, T; W^{1,\infty}(r_0, \infty))$ for every $(r_0, \tau) \in Q$.*

Proof. Fix $1 < r_0 < r_1 < r_2$ and $0 < \tau < T$. By Proposition 4.2,

$$u(r,t) \leq K_0\left\{t^{-1}N_0^2(t) + N_0^{m+1}(t)\right\}^{1/(m+1)}$$

for all $r \geq r_1$ and $0 < t < T$, where K_0 is some constant which depends only on m, n, r_0, and r_1, and

$$N_0(t) := \sup\left\{\int_{r_0}^\infty x^{n-1}u(x,s)\,dx : 0 < s \leq t\right\}.$$

Hence,

$$v(r,t) \leq K_0\left\{\tau^{-1}TN_1^2(t) + N_1^{m+1}(t)\right\}^{1/(m+1)} \tag{5.20}$$

for all $r \geq r_1$ and $\tau \leq t < T$, where

$$N_1(t) := \sup\left\{(T-s)^{1/(m-1)}\int_{r_0}^\infty x^{n-1}u(x,s)\,dx : 0 < s \leq t\right\}.$$

Simultaneously, by Proposition 4.1,

$$\left|u^{m-1}(x,t) - u^{m-1}(y,t)\right| \leq K_1\left\{t^{-1}M_0^{m-1}(t) + M_0^{2(m-1)}(t)\right\}^{1/2}|x-y|$$

for all $x \geq r_2$, $y \geq r_2$, and $0 < t < T$, where K_1 is some constant which depends only on m, n, r_1, and r_2 and

$$M_0(t) := \sup\{u(r,s) : r \geq r_1, 0 < s \leq t\}.$$

Hence,

$$\left|v^{m-1}(x,t) - v^{m-1}(y,t)\right| \leq K_1 \left\{\tau^{-1}TM_1^{m-1}(t) + M_1^{2(m-1)}(t)\right\}^{1/2} |x-y| \tag{5.21}$$

for all $x \geq r_2$, $y \geq r_2$, and $\tau \leq t < T$, where

$$M_1(t) := \sup\{v(r,s) : r \geq r_1, 0 < s \leq t\}.$$

Now, since $L(1) = \mathcal{C} < \infty$, N_1 is bounded on $(0,T)$ by Lemma 5.6. Thus, from (5.20), we deduce that $v^{m-1} \in L^\infty((r_1,\infty)\times(\tau,T))$. In turn, because u itself is bounded in $(1,\infty)\times(0,\tau]$, this means that M_1 is bounded on $(0,T)$. Subsequently from (5.21) we deduce that $v^{m-1} \in L^\infty(\tau,T;W^{1,\infty}(r_2,\infty))$. These deductions provide the lemma, in the light of the arbitrariness of r_0, r_1, and r_2. □

LEMMA 5.11. *Let u_1 and u_2 be any two weak solutions of Problem* 1.1 *with corresponding functions $u_{0,i}$, f_i, ℓ_i, and z_i, and parameters $\mathcal{C}_i$ for $i = 1,2$. Suppose that* (1.14) *holds. Then $\ell_1(r) \leq \ell_2(r)$ for all $r > 1$.*

Proof. We borrow ideas from [14, 26]. Since when $\mathcal{C}_2 = \infty$ the result is immediate from Lemma 5.9, without loss of generality we may assume that $\mathcal{C}_2 < \infty$. Define $u_{0,3}(r) := \max\{u_{0,1}(r), u_{0,2}(r)\}$ for $r > 1$, and $f_3(t) := \max\{f_1(t), f_2(t)\}$ for $0 < t < T$. Denote by u_3 the solution of Problem 1.1 with initial data function $u_{0,3}$ and boundary data function f_3. Let ℓ_3, $\mathcal{C}_3$, z_3, and v_3 be defined in the obvious way. Note that $\mathcal{C}_3 = \mathcal{C}_2$ by hypothesis (1.14). Now, by the comparison principle for solutions of Problem 1.1 [18, 19], $u_i(r,t) \leq u_3(r,t)$ for all $(r,t) \in Q$ and $i = 1,2$. Hence, $\ell_1(r) \leq \ell_3(r)$ for all $r > 1$. Subsequently, it suffices to show that

$$\ell_3(r) \leq \ell_2(r) \quad \text{for all } r > 1. \tag{5.22}$$

To do this, note that Lemma 5.10 can be applied to both u_2 and u_3 since $\mathcal{C}_2 = \mathcal{C}_3 < \infty$. Applying an argument in [2, 14, 26], this lemma means that for every $(r_0,\tau) \in Q$ each of the functions $v_i(r,t)$ with $i = 2,3$ is uniformly Hölder continuous with respect to r with exponent

$$\iota := \min\{1, 1/(m-1)\}$$

in $[r_0,\infty)\times[\tau,T)$. Subsequently, for fixed $r_0 > 1$ and $0 < \tau < T$, the function $v_3 - v_2$ is uniformly Hölder continuous with respect to r with exponent ι in $[r_0,\infty)\times[\tau,T)$. In particular, this means that

$$(v_3 - v_2)(x,t) \geq (v_3 - v_2)(r_0,t) - K(x-r_0)^\iota \tag{5.23}$$

for all $x \geq r_0$ and $\tau \leq t < T$, for some positive constant K. Now, by (2.1), the nonnegativity of $v_3 - v_2$, and (2.4),

$$\begin{aligned}(T-t)^{1/(m-1)}(z_3 - z_2)(1,t) &= \int_1^\infty \kappa(x,1)(v_3 - v_2)(x,t)\,dx \\ &\geq \int_{r_0}^r \kappa(x,1)(v_3 - v_2)(x,t)\,dx \\ &\geq \kappa(r_0,1)\int_{r_0}^r (v_3 - v_2)(x,t)\,dx\end{aligned} \tag{5.24}$$

for any $r \geq r_0$ and $\tau \leq t < T$. Subsequently, choosing $r := r_0 + K^{-1/\iota}(v_3 - v_2)^{1/\iota}(r_0, t)$ and substituting (5.23) in (5.24), we deduce that

$$\left\{\frac{\iota+1}{\iota}\frac{K}{\kappa(r_0,1)}\,(T-t)^{1/(m-1)}\,(z_3-z_2)(1,t)\right\}^{\iota/(\iota+1)} \geq (v_3-v_2)(r_0,t)$$

for all $\tau \leq t < T$. Passing to the limit $t \uparrow T$, using (5.2), (1.12), and (1.14), this inequality implies $\ell_3(r_0) \leq \ell_2(r_0)$, whence, because, $r_0 > 1$ was arbitrary, (5.22) is obtained. □

Lemmas 5.3 and 5.11 together constitute Theorem 1.8.

5.4. The proof of Theorem 1.7. We build up the proof of Theorem 1.7 from the analysis so far in five steps. We make each step the content of a lemma. The first two are concerned with the properties of L and the others with the properties of ℓ.

LEMMA 5.12. *If $\mathcal{C} < \infty$, then* (i) $L(r_0) \geq r_0^{2m/(m-1)}\mathcal{A}(\omega/r_0)$ *for all* $1 \leq r_0 \leq \omega$, (ii) $L(r_0) > L(r)$ *for all* $1 \leq r_0 < r \leq \omega$, *and* (iii) $L(r) = 0$ *for all* $r > \omega$.

Proof. Fix $1 \leq r_0 \leq \omega$ and define r_1 by (5.18). By Lemma 5.8, $z(r,T) < \infty$ for all $r > r_1$. Thus, by Lemma 5.6 and Theorem 1.5, $r_1 \geq \omega$. Rewriting $r_1 \geq \omega$ provides part (i). Next, substituting (5.5) in (5.17), there holds

$$L(r) \leq (m-1)r_1^{2m/(m-1)}\rho^m(r/r_1) \qquad \text{for any } r_0 \leq r \leq r_1,$$

where ρ is the function from Lemma 3.2. Hence, because ρ is strictly decreasing on $(0,1]$,

$$L(r) < (m-1)r_1^{2m/(m-1)}\rho^m(r_0/r_1) = r_0^{2m/(m-1)}\mathcal{A}(r_1/r_0) = L(r_0)$$

for any $r_0 < r \leq r_1$. Thus, since $r_1 \geq \omega$, part (ii) is proven. Part (iii) is a simple consequence of (5.2) and Lemma 5.7. □

LEMMA 5.13. *If $\mathcal{C} < \infty$ then $L \in C(1,\infty) \cap W^{1,\infty}_{\mathrm{loc}}(1,\infty)$.*

Proof. Integrating (2.7) we determine that

$$z(r_1,t)\int_{r_0}^{r_2} x^{1-n}\,dx \leq z(r_0,t)\int_{r_1}^{r_2} x^{1-n}\,dx + z(r_2,t)\int_{r_0}^{r_1} x^{1-n}\,dx \tag{5.25}$$

for any $1 \leq r_0 < r_1 < r_2$ and $0 < t < T$. Hence, multiplying by $(T-t)^{1/(m-1)}$ and then letting $t \uparrow T$, (5.25) holds with $L(r_i)$ in lieu of $z(r_i,t)$ for $i \in \{0,1,2\}$. This means that subject to a transformation of its argument, the function L is convex in $[1,\infty)$. The stated regularity follows from this. □

LEMMA 5.14. *Let ρ be the function from Lemma* 3.2. *If $c > 1$ is such that for some $0 < \tau < T$ there holds $f(t) \leq (T-t)^{-1/(m-1)}c^{2/(m-1)}\rho(1/c)$ for almost all $\tau < t < T$, then v is bounded in Q, and $\ell(r) \leq c^{2m/(m-1)}\rho(\min\{r/c,1\})$ for all $r > 1$.*

Proof. Let $U(\cdot,\cdot;a,T)$ be the solution of (1.1) defined in Proposition 3.1. By hypothesis, $U(1,t;c,T) \geq f(t)$ for almost all $\tau < t < T$. Consequently, we can choose an a so large that $U(1,t;a,T) \geq f(t)$ for almost all $0 < t < T$ and $U(r,0;a,T) \geq u_0(r)$ for almost all $r > 1$. The comparison principle [18, 19] for solutions of Problem 1.1 then says that $U(r,t;a,T) \geq u(r,t)$ for all $(r,t) \in Q$. This gives the first conclusion of the lemma. The second one is a corollary of Lemma 5.11 taking $u_1 = u$ and $u_2 = U(\cdot,\cdot;c,T)$. □

LEMMA 5.15. *Let ρ be the function from Lemma* 3.2. *If $\mathcal{C} < \infty$ then* (i) $\ell(r_0) \geq \omega^{2/(m-1)}\rho(r_0/\omega)$ *for all* $1 < r_0 \leq \omega$, (ii) $\ell(r_0) > \ell(r)$ *for all* $1 < r_0 < r \leq \omega$, *and* (iii) $\ell(r) = 0$ *for all* $r > \omega$.

Proof. Fix $r_0 > 1$, and note that $\ell(r_0) < \infty$ by Lemma 5.10. Applying the scaling trick used to obtain Lemma 5.8 from Lemma 5.4. Lemma 5.14 yields

$$\ell(r) \leq r_1^{2/(m-1)}\rho(\min\{r/r_1, 1\}) \quad \text{for all } r > r_0, \tag{5.26}$$

where $r_1 \geq r_0$ is such that

$$\ell(r_0) = r_1^{2/(m-1)}\rho(r_0/r_1). \tag{5.27}$$

In particular this means that $\ell(r_1) = 0$. By (5.19) this in turn implies that $L(r_1) = 0$. Therefore, by Lemma 5.12 part (ii), $r_1 \geq \omega$. Substituting this estimate in (5.27) gives part (i). On the other hand, in view of the monotonicity of ρ, (5.26) and (5.27) imply that $\ell(r) \leq \ell(r_0)$ for all $r > r_0$ with equality only if $\min\{r/r_1, 1\} = r_0/r_1$, i.e., $r_1 = r_0$. However, by (5.27), the latter is the case if and only if $\ell(r_0) = 0$. Thus, we deduce that $\ell(r) < \ell(r_0)$ for all $r > r_0$ such that $\ell(r_0) > 0$. Recalling part (i), this yields part (ii). Part (iii) follows immediately from the definitions of ℓ and ω. □

The assertions of Theorem 1.7 concerning ℓ and L can be distilled from Lemmas 5.9, 5.10, 5.12, 5.13, and 5.15. To prove Theorem 1.7 it therefore remains to show that when f is nondecreasing, $u(\cdot, t) \to \infty$ as $t \uparrow T$ uniformly on $(1, r)$ for every $1 < r < \omega$. However, since for this purpose, without loss of generality we may suppose that $u_0 \equiv 0$, this follows from Lemma 5.1.

5.5. The proof of Theorem 1.9. To prove Theorem 1.9 we shall use one additional lemma.

LEMMA 5.16. *If $a > 1$ is such that $\mathcal{A}(a) < \mathcal{C}$ then $z(r, T) = \infty$ for all $1 < r < a$ for which $r^{2m/(m-1)}\mathcal{A}(a/r) > 2^{-1/(m-1)}\mathcal{A}(a)$.*

Proof. The proof of this lemma is rather similar to that of Lemma 5.5. By definition, there exists a sequence of values $\{t_i\}_{i=1}^{\infty} \subset (0, T)$ such that $t_i \to T$ as $i \to \infty$, and $(T - t_i)^{1/(m-1)}z(1, t_i) > \mathcal{A}(a)$ for any $i \geq 1$. Fix i and set $\tau_i := 2T - t_i$. Then $z(1, t) \geq z(1, t_i) > (T - t_i)^{-1/(m-1)}\mathcal{A}(a) = Z_w(1, T; a, \tau_i) \geq Z_w(1, t; a, \tau_i)$ for all $t_i < t < T$. Proposition 2.2 subsequently states that $z(r, t) \geq Z_w(r, t; a, \tau_i) - \max\{Z_w(1, s; a, \tau_i) : 0 \leq s \leq t_i\} = Z_w(r, t; a, \tau_i) - Z_w(1, t_i; a, \tau_i)$ for all $(r, t) \in Q$. Hence, $z(r, T) \geq Z_w(r, T; a, \tau_i) - Z_w(1, t_i; a, \tau_i) = (T - t_i)^{-1/(m-1)}\{r^{2m/(m-1)}\mathcal{A}(a/r) - 2^{-1/(m-1)}\mathcal{A}(a)\}$ for all $1 < r < a$. Letting $i \to \infty$ yields the result. □

Now, Lemmas 5.4, 5.5, and 5.16 imply that $z(r, T) < \infty$ for all $r > \alpha(\mathcal{C})$, $z(r, T) = \infty$ for all $1 < r < \beta(\mathcal{C})$, and $z(r, T) = \infty$ for all $1 < r < \gamma(\mathcal{C})$, respectively. While, by Lemmas 5.7 and 5.6, $z(r, T) < \infty$ for all $r > \omega$, and $z(r, T) = \infty$ for all $1 < r < \omega^*$. Combining these conclusions with Theorem 1.5 yields

$$\max\{\beta(\mathcal{C}), \gamma(\mathcal{C})\} \leq \omega \leq \alpha(\mathcal{C}). \tag{5.28}$$

Theorem 1.9 is subsequently an immediate consequence of Lemma 5.2.

5.6. On the size of ℓ and L. Our last result on Problem 1.1 provides some quantitative information on the way blow-up occurs.

THEOREM 5.17. *Let u be a weak solution of Problem 1.1 for which $0 < \mathcal{C} < \infty$ where $\mathcal{C}$ is defined by (1.10). Let ℓ and L be defined by (1.9) and (1.12). Also, let ρ be the function from Lemma 3.2 and let α be the inverse of the function $\mathcal{A}$ defined by (5.5). Then*

$$\omega \leq \alpha(\mathcal{C}), \tag{5.29}$$

$$\mathcal{A}(\omega/r) \leq r^{-2m/(m-1)}L(r) \leq \mathcal{A}(\alpha(\mathcal{C})/r) \quad \textit{for all } 1 \leq r \leq \omega, \tag{5.30}$$

and

$$\ell(r) \geq \omega^{2/(m-1)}\rho(r/\omega) \quad \textit{for all } 1 < r \leq \omega. \tag{5.31}$$

(a) *If furthermore*

$$(T-t)^{1/(m-1)} \int_0^t f^m(s)\,ds \to \mathcal{C} \quad \textit{as } t \uparrow T,$$

then (5.29) *and* (5.30) *hold with equality.*

(b) *If furthermore*

$$\operatorname{ess\,sup}\{(T-s)^{1/(m-1)} f(s) : t < s < T\} \to (m-1)^{-1/m}\mathcal{C}^{1/m} \quad \textit{as } t \uparrow T,$$

then (5.29)–(5.31) *hold with equality.*

Proof. The main assertions have virtually been established already. The inequality (5.29), the left-hand inequality in (5.30), and (5.31) are restatements of (5.28), Lemma 5.12 part (i), and Lemma 5.15 part (i). The remaining inequality in (5.30) may be obtained by letting $c \downarrow \alpha(\mathcal{C})$ in Lemma 5.4. Moreover, under the additional hypothesis of part (a), applying Lemma 5.3 with $z_1 = Z_w(\cdot, \cdot; \alpha(\mathcal{C}), T)$ and $z_2 = z$ yields

$$L(r) \geq r^{2m/(m-1)}\mathcal{A}(\alpha(\mathcal{C})/r) > 0 \quad \text{for all } 1 \leq r < \alpha(\mathcal{C}). \tag{5.32}$$

Hence, by Lemma 5.12 part (iii), equality in (5.29) is obtained, and part (a) is proven. Let us therefore assume that the hypothesis of part (b) holds. In this case, Lemma 5.14 gives

$$\ell(r) \leq \{\alpha(\mathcal{C})\}^{2/(m-1)}\,\rho(r/\alpha(\mathcal{C})) \quad \text{for all } 1 < r \leq \alpha(\mathcal{C}). \tag{5.33}$$

Simultaneously, integrating (2.7), there holds

$$\begin{aligned} z(r_0,t)\int_{r_1}^{r_2} x^{1-n}\,dx + z(r_2,t)\int_{r_0}^{r_1} x^{1-n}\,dx \\ = z(r_1,t)\int_{r_0}^{r_2} x^{1-n}\,dx + \int_{r_0}^{r_1}\int_{r_1}^{r_2}\int_x^y x^{1-n}y^{1-n}\eta^{n-1}u(\eta,t)\,d\eta\,dy\,dx \end{aligned}$$

for any $1 \leq r_0 < r_1 < r_2$ and $0 < t < T$. Taking $r_0 = 1$, $r_1 = r$, and $r_2 = \alpha(\mathcal{C})$, multiplying by $(T-t)^{1/(m-1)}$, letting $t \uparrow T$ in an appropriate fashion, using (5.2) and (1.10), and employing Fatou's lemma which is justified by Lemma 5.14, this gives

$$\begin{aligned} \mathcal{C}\int_r^{\alpha(\mathcal{C})} x^{1-n}\,dx \leq L(r)\int_1^{\alpha(\mathcal{C})} x^{1-n}\,dx \\ + \int_1^r\int_r^{\alpha(\mathcal{C})}\int_x^y x^{1-n}y^{1-n}\eta^{n-1}\ell(\eta)\,d\eta\,dy\,dx \end{aligned} \tag{5.34}$$

for all $1 < r < \alpha(\mathcal{C})$. Repeating the above calculation with the solution $U(\cdot, \cdot; \alpha(\mathcal{C}), T)$ from Proposition 3.1 instead of u, one obtains

$$\begin{aligned} \mathcal{C}\int_r^{\alpha(\mathcal{C})} x^{1-n}\,dx = r^{2m/(m-1)}\mathcal{A}(\alpha(\mathcal{C})/r)\int_1^{\alpha(\mathcal{C})} x^{1-n}\,dx \\ + \int_1^r\int_r^{\alpha(\mathcal{C})}\int_x^y x^{1-n}y^{1-n}\eta^{n-1}\{\alpha(\mathcal{C})\}^{2/(m-1)}\rho(\eta/\alpha(\mathcal{C}))\,d\eta\,dy\,dx \end{aligned} \tag{5.35}$$

for all $1 < r < \alpha(\mathcal{C})$. Combining (5.33), (5.34), and (5.35) yields (5.32). This once again implies equality in (5.29) and (5.30). The outstanding conclusion subsequently follows from (5.31) and (5.33). $\square$

In the case $n = 1$, the functions $\mathcal{A}$ and α can be computed explicitly from the function ρ given by (3.10).

6. The Cauchy–Neumann problem.

6.1. Notation. We shall retain the conventions used in the study of the Cauchy–Dirichlet problem in the previous section. We let u denote a given weak solution of Problem 1.2 with initial data function u_0 and boundary data function g. We have the following result in lieu of Lemma 5.1.

LEMMA 6.1. *Suppose that g is nondecreasing and $u_0 \equiv 0$. Then $u(r,t) \geq u(r_0,t_0)$ for all $1 < r \leq r_0$ and $0 < t_0 \leq t < T$.*

Proof. We use a technique applied in [10]. Pick $0 < \tau < T$ and $0 < \varepsilon < T - \tau$. Define $u_{0,\varepsilon}(r) := u(r,\varepsilon)$ for $r > 1$, and $g_\varepsilon(t) := g(t+\varepsilon)$ for $0 < t < \tau$. Consider Problem 1.2 restricted to the domain Q_τ with initial data $u_{0,\varepsilon}$ and boundary data g_ε. By the general theory, this problem admits a unique weak solution $u_\varepsilon(r,t) = u(r,t+\varepsilon)$. Furthermore, since $u_{0,\varepsilon}(r) \geq 0 = u_0(r)$ for all $r > 1$ and $g_\varepsilon(t) = g(t+\varepsilon) \geq g(t)$ for almost all $0 < t < \tau$, there holds $u_\varepsilon(r,t) \geq u(r,t)$ for all $(r,t) \in Q_\tau$. Thus $u(r,t+\varepsilon) \geq u(r,t)$ for all $(r,t) \in Q_\tau$. In view of the arbitrariness of ε and τ, this yields the monotonicity of u with respect to t. In fact, since u is a classical solution of (1.1) at any point in Q where it is positive, it yields $(\partial u/\partial t)(r,t) \geq 0$ for all $(r,t) \in Q$ such that $u(r,t) > 0$. Hence, by (1.1), $(\partial(r^{n-1}\partial u^m/\partial r)/\partial r)(r,t) \geq 0$ for all such points. Recalling that $\partial u^m/\partial r \in C(Q)$, noting that as a consequence of the nonnegativity of u there holds $(\partial u^m/\partial r)(r,t) = 0$ for all $(r,t) \in Q$ such that $u(r,t) = 0$, and recalling that $u(r,t) = 0$ for all $r > \zeta(t)$ and $0 < t < T$, this implies that $(\partial u^m/\partial r)(r,t) \leq 0$ for all $(r,t) \in Q$. Hence, we also have the monotonicity of u with respect to r. $\square$

We employ the notation (1.5), (1.6), (1.9), and (5.3). We define $\mathcal{C}$ by (1.11), L by (1.16), and z by (2.18) and (2.19) and for completeness set

$$z(r,T) := \limsup_{t \uparrow T} z(r,t).$$

We let

$$Z(r,t;a,\tau) := \int_r^\infty x^{n-1} U(x,t;a,\tau)\,dx$$

for any $(r,t) \in [1,\infty) \times [0,T)$ for one of the self-similar solutions U of the equation, and add a subscript w if the self-similar solution is that of waiting-time type and a subscript s if it is that of instantaneous point-source type. Performing calculations analogous to those we have carried in section 5 we obtain

$$Z_w(r,t;a,\tau) = \begin{cases} (\tau - t)^{-1/(m-1)} r^{\{(m-1)n+2\}/(m-1)} \mathcal{A}(a/r) & \text{if } r < a, \\ 0 & \text{otherwise} \end{cases}$$

for all $(r,t) \in [1,\infty) \times [0,T)$, where now

$$\mathcal{A}(c) := c^{\{(m-1)n+2\}/(m-1)} \int_{1/c}^1 x^{n-1} \rho(x)\,dx \tag{6.1}$$

for $c \geq 1$, with ρ constructed in Lemma 3.2. Similarly

$$Z_s(r,t;a,\tau) \\ = \begin{cases} (t-\tau)^{-1/(m-1)} r^{\{(m-1)n+2\}/(m-1)} \mathcal{B}(\xi(t)/r) & \text{if } t > \tau \text{ and } r < \xi(t), \\ 0 & \text{otherwise} \end{cases}$$

for all $(r,t) \in \overline{Q}$, where

$$\mathcal{B}(c) := c^{\{(m-1)n+2\}/(m-1)} \int_{1/c}^{1} x^{n-1} \rho(x)\, dx$$

for $c \geq 1$, with ξ and ρ defined by (3.12) and (3.13). The functions $\mathcal{A}$ and $\mathcal{B}$ are unbounded, strictly increasing, and continuously differentiable on $[1,\infty)$ and are such that $\mathcal{A}(1) = \mathcal{B}(1) = 0$. We shall denote by α and β the inverse of the functions $\mathcal{A}$ and $\mathcal{B}$, respectively. Finally, define the function $\gamma \in C([0,\infty))$ with $\gamma(0) = 1$ by

$$\{\gamma(\mathcal{C})\}^{\{(m-1)n+2\}/(m-1)} \mathcal{A}(\alpha(\mathcal{C})/\gamma(\mathcal{C})) = 2^{-1/(m-1)}\mathcal{C}.$$

The proof of the following is similar to the proof of Lemma 5.2 and we omit it.

LEMMA 6.2. *There are positive constants μ and ν such that* $\max\{\beta(\mathcal{C}), \gamma(\mathcal{C})\} \geq 1 + \mu\{\sigma(\mathcal{C})\}^{(m-1)/(m+1)}$ *and* $\alpha(\mathcal{C}) \leq 1 + \nu\{\sigma(\mathcal{C})\}^{(m-1)/(m+1)}$ *for all* $\mathcal{C} \geq 0$*, where* σ *is defined by* (1.17).

6.2. The proof of Theorem 1.10. The proofs of the next two lemmas are almost identical to those of their counterparts Lemmas 5.4 and 5.5 in the previous section, and will be omitted. The third lemma below is a consequence of the definitions of Ω^* and z.

LEMMA 6.3. *If* $c > 1$ *is such that* $\mathcal{A}(c) > \mathcal{C}$ *then*

$$L(r) \leq r^{\{(m-1)n+2\}/(m-1)} \mathcal{A}(\max\{c/r, 1\}) \quad \text{for all } r \geq 1$$

and $z(r,T) < \infty$ *for all* $r \geq c$.

LEMMA 6.4. *If* $c > 1$ *is such that* $\mathcal{B}(c) < \mathcal{C}$*, then* $z(r,T) = \infty$ *for all* $1 < r < c$.

LEMMA 6.5. *There holds* $z(r,T) = \infty$ *for some* $r > 1$ *if and only if* $r \in \Omega^*$.

With these lemmas we may complete the proof of Theorem 1.10. Suppose that $\mathcal{C} < \infty$, then by Lemma 6.3, $z(r,T) < \infty$ for all $r > \alpha(\mathcal{C})$. Hence, by Lemma 6.5, $\omega^* < \infty$. On the other hand if $\mathcal{C} = \infty$, then by Lemma 6.4, $z(r,T) = \infty$ for all $r > 1$. Whence, by Lemma 6.5, $\omega^* = \infty$. The desired result consequently follows from Theorem 1.5.

6.3. The proof of Theorem 1.12. Comparing the Cauchy–Neumann and Cauchy–Dirichlet problems, Theorem 1.12 is the counterpart of Theorem 1.8. The last-mentioned theorem was proven in two steps, viz., Lemma 5.3 and Lemma 5.11. The analogous lemmas for the Cauchy–Neumann problem can be formulated and proven without any extra difficulty. Details will not be presented.

6.4. The proof of Theorem 1.11. The proof of Theorem 1.11 can be completed similarly to the proof of Theorem 1.7 once we have established Lemmas 6.7–6.9 below. Therefore, we shall not dwell on any details beyond the proofs of these lemmas. As a means to obtaining Lemma 6.7, we use Lemma 6.6 below. This plays the role for Problem 1.2 which Lemma 5.8 does for Problem 1.1. However, for technical reasons, we defer the proof of this lemma until the next section.

LEMMA 6.6. *Suppose that $L(r_0) < \infty$ for some $r_0 > 1$. Then*

$$L(r) \leq r^{\{(m-1)n+2\}/(m-1)} \mathcal{A}(\max\{r_1/r, 1\}) \quad \text{for all } r \geq r_0,$$

and $z(r,T) < \infty$ for all $r > r_1$, where $r_1 := r_0 \alpha(r_0^{-\{(m-1)n+2\}/(m-1)} L(r_0))$.

LEMMA 6.7. *If $\mathcal{C} < \infty$ then* (i) *$L(r) \geq r^{\{(m-1)n+2\}/(m-1)} \mathcal{A}(\omega/r)$ for all $1 \leq r \leq \omega$,* (ii) *$L(r_0) > L(r)$ for all $1 \leq r_0 < r \leq \omega$, and* (iii) *$L(r) = 0$ for all $r > \omega$. On the other hand if $\mathcal{C} = \infty$, then $L(r) = \infty$ for all $r > 1$.*

Proof. This lemma follows from Lemma 6.6 in the same way as Lemmas 5.9 and 5.12 follow from Lemma 5.8. □

LEMMA 6.8. *Let ρ be the function from Lemma* 3.2. *If $\mathcal{C} < \infty$ then* (i) *$v^{m-1} \in L^\infty(\tau, T; W^{1,\infty}(r_1, \infty))$ for every $(r_1, \tau) \in Q$,* (ii) *$\ell(r) \geq \omega^{2/(m-1)} \rho(r/\omega)$ for all $1 < r \leq \omega$,* (iii) *$\ell(r_1) > \ell(r)$ for all $1 < r_1 < r \leq \omega$, and* (iv) *$\ell(r) = 0$ for all $r > \omega$. On the other hand if $\mathcal{C} = \infty$, then $\ell(r) = \infty$ for all $r > 1$.*

Proof. Fix $r_0 > 1$. Then the function $\widetilde{u}$ defined by (2.9) is a solution of Problem 1.1 with correspondingly scaled initial data and boundary data given by (2.10). Denote, associated with $\widetilde{u}$, the parameter $\widetilde{\mathcal{C}}$ defined by (1.10). By Theorem 1.10, u is effectively localized if and only if $\mathcal{C} < \infty$. However, by Theorem 1.6, $\widetilde{u}$ is effectively localized if and only if $\widetilde{\mathcal{C}} < \infty$. Consequently, $\widetilde{\mathcal{C}} < \infty$ if and only if $\mathcal{C} < \infty$. The present lemma may now be deduced by applying Lemmas 5.9, 5.10, and 5.15 to $\widetilde{u}$, transferring the conclusions to u, and finally, letting $r_0 \downarrow 1$. □

LEMMA 6.9. *If $\mathcal{C} < \infty$ then $L \in C(1,\infty) \cap W^{1,\infty}_{\mathrm{loc}}(1,\infty)$.*

Proof. By (2.18) there holds

$$z(r_0, t) = z(r,t) + \int_{r_0}^{r} x^{n-1} u(x,t)\, dx \quad \text{for all } 1 < r_0 < r$$

and $0 < t < T$. Multiplying this inequality by $(T-t)^{1/(m-1)}$, letting $t \uparrow T$, and employing Fatou's lemma which is justified by the previous lemma give

$$L(r_0) \leq L(r) + \int_{r_0}^{r} x^{n-1} \ell(x)\, dx \quad \text{for all } 1 < r_0 < r.$$

In light of Lemmas 6.7 and 6.8, this provides the stated regularity. □

6.5. The proof of Theorem 1.13. With regard to Theorem 1.13 we employ one more lemma. Its proof is almost a facsimile of the proof of Lemma 5.16 and is not included.

LEMMA 6.10. *If $a > 1$ is such that $\mathcal{A}(a) < \mathcal{C}$ then $z(r,T) = \infty$ for all $1 < r < a$ for which $r^{\{(m-1)n+2\}/(m-1)} \mathcal{A}(a/r) > 2^{-1/(m-1)} \mathcal{A}(a)$.*

Now, suppose that $\mathcal{C} < \infty$. Then by Lemma 6.3, there holds $z(r,T) < \infty$ for all $r > \alpha(\mathcal{C})$, while by Lemmas 6.4 and 6.10, there holds $z(r,T) = \infty$ for all $1 < r < \max\{\beta(\mathcal{C}), \gamma(\mathcal{C})\}$. Recalling Lemma 6.5 and Theorem 1.5, this gives (5.28). Theorem 1.13 follows from Lemma 6.2.

6.6. On the size of ℓ and L. Analogous to Theorem 5.17 for the Cauchy–Dirichlet problem, the next result can be obtained for Problem 1.2, where, in the case $n = 1$, the functions $\mathcal{A}$ and α can be computed explicitly from the function ρ given by (3.10).

THEOREM 6.11. *Let u be a weak solution of Problem* 1.2 *for which $0 < \mathcal{C} < \infty$ where $\mathcal{C}$ is defined by* (1.11). *Let ℓ and L be defined by* (1.9) *and* (1.16). *Also, let ρ*

be the function from Lemma 3.2 *and let α be the inverse of the function $\mathcal{A}$ defined by* (6.1)*. Then*

$$\omega \leq \alpha(\mathcal{C}), \tag{6.2}$$

$$\mathcal{A}(\omega/r) \leq r^{-\{(m-1)n+2\}/(m-1)}L(r) \leq \mathcal{A}(\alpha(\mathcal{C})/r) \quad \textit{for all } 1 < r \leq \omega, \tag{6.3}$$

and

$$\ell(r) \geq \omega^{2/(m-1)}\rho(r/\omega) \quad \textit{for all } 1 < r \leq \omega. \tag{6.4}$$

(a) *If furthermore*

$$(T-t)^{1/(m-1)}\int_0^t g(s)\,ds \to \mathcal{C} \qquad \textit{as } t \uparrow T,$$

then (6.2) *and* (6.3) *hold with equality.*

(b) *If furthermore*

$$\operatorname{ess\,sup}\{(T-s)^{m/(m-1)}g(s) : t < s < T\} \to (m-1)^{-1}\mathcal{C} \qquad \textit{as } t \uparrow T,$$

then (6.2)–(6.4) *hold with equality.*

7. The remaining details. Without the assumption that g is nonnegative and bounded on every interval $(0,\tau)$ with $0 < \tau < T$, it is a priori unclear whether the Cauchy–Neumann problem (Problem 1.2) admits a nonnegative solution. Notwithstanding, if we suppose that the problem has a nonnegative solution which can be constructed as the limit of a sequence of positive solutions of (1.1), a number of our results can still be shown to hold. In particular, under the assumption (2.23), Lemma 2.3 remains true. Consequently, mutatis mutandi, it is also possible to obtain Proposition 2.4, and Lemmas 6.3 and 6.5. It follows that with $\mathcal{C}$ defined by (1.11), as long as (2.23) holds, even if g changes sign and is unbounded, the criterion $\mathcal{C} < \infty$ remains sufficient for the localization, effective localization, and metastable localization of a nonnegative solution of Problem 1.2 which can be constructed as the limit of a sequence of positive solutions of (1.1).

It is this observation which enables us to prove Lemma 6.6.

Proof of Lemma 6.6. We apply the scaling technique used in the proof of Lemmas 2.1 and 5.8. The function $\widetilde{u}$ given by (2.9) is a weak solution of Problem 1.2 with correspondingly scaled initial data, and with boundary data

$$\widetilde{g}(t) := r_0^{-(m+1)/(m-1)}\frac{\partial u^m}{\partial r}(r_0,t), \tag{7.1}$$

for which the parameter defined by (1.11) takes the value $r_0^{-\{(m-1)n+2\}/(m-1)}L(r_0)$. Applying Lemma 6.3 to the rescaled problem, which can now be justified with the help of the above remarks, gives that which is required. □

As a consequence of Lemma 6.6, for any solution u of Problem 1.2 which can be constructed as the limit of a sequence of positive solutions of (1.1) and whose boundary data satisfy (2.23), with the variables otherwise as defined in Theorem 1.11, we can still obtain Theorems 1.11 and 6.11.

Appendix. For the one-dimensional case, under appropriate assumptions on u_0 and f, it was shown in [15] that Problem 1.1 displays localization if and only if the parameter $\mathcal{C}^*$ defined by (1.8) is finite. On the other hand, in [11] it was shown that this problem displays localization if and only if the parameter $\mathcal{C}$ defined by (1.10) is finite. This naturally raises the question if there is a discrepancy in these criteria, and if not, which is to be preferred. In both [15] and [11] it was hypothesized that f is continuous. The key to resolving the question of the relation between the criteria is the observation that in [15] the additional hypothesis that f is nondecreasing was imposed. In this appendix we show that with this additional hypothesis the two criteria are equivalent. Thus, also being applicable when f is not monotonic, the criterion of Cortazar and Elgueta [11] constitutes the superior result.

Let $f \in C(0,T)$ be nondecreasing. Fix $0 < t < T$ and set

$$\tau := \{t + (m-1)T\}/m. \tag{A.1}$$

Then, noting that $t < \tau < T$, there holds

$$\begin{aligned} f(t) &= (\tau - t)^{-1/m} \left\{ \int_t^\tau f^m(t)\, ds \right\}^{1/m} \\ &\le (\tau - t)^{-1/m} \left\{ \int_t^\tau f^m(s)\, ds \right\}^{1/m} \\ &\le (\tau - t)^{-1/m} \left\{ \int_0^\tau f^m(s)\, ds \right\}^{1/m}. \end{aligned}$$

Subsequently multiplying both sides of this last inequality by $(T-t)^{1/(m-1)}$ and using (A.1) we deduce

$$(T-t)^{1/(m-1)} f(t) \le m^{1/(m-1)} (m-1)^{-1/m} \left\{ (T-\tau)^{1/(m-1)} \int_0^\tau f^m(s)\, ds \right\}^{1/m}.$$

This means that

$$\limsup_{t \uparrow T} \, (T-t)^{1/(m-1)} \, f(t) \le m^{1/(m-1)} (m-1)^{-1/m} \mathcal{C}^{1/m}. \tag{A.2}$$

However, if (A.2) holds, by Lemma A2 in [15] with $\phi(t) = (T-t)^{-1/(m-1)}$ we have

$$\mathcal{C}^* \le m \left\{ m^{1/(m-1)} (m-1)^{-1/m} \mathcal{C}^{1/m} \right\}^{m-1}. \tag{A.3}$$

On the other hand, by Lemma 5 in [15] there holds

$$\mathcal{C} \le (m-1)^{-1/(m-1)} \left\{ \mathcal{C}^* \right\}^{m/(m-1)}. \tag{A.4}$$

Combining (A.3) and (A.4) yields

$$(m-1)^{1/m} \mathcal{C}^{(m-1)/m} \le \mathcal{C}^* \le m^2 (m-1)^{(1-m)/m} \mathcal{C}^{(m-1)/m}.$$

This shows that under the hypothesis that f is nondecreasing and continuous, the criteria of [15] and [11] are equivalent.

REFERENCES

[1] L. Alvarez and J. I. Diaz, *Sufficient and necessary initial mass conditions for the existence of a waiting time in nonlinear-convection processes*, J. Math. Anal. Appl., 155 (1991), pp. 378–392.

[2] D. G. Aronson, *Regularity properties of flows through porous media*, SIAM J. Appl. Math., 17 (1969), pp. 461–467.

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

[4] D. G. Aronson, *Regularity of flows in porous media: A survey*, in Nonlinear Diffusion Equations and Their Equilibrium States I, W.-M. Ni, L. A. Peletier, and J. Serrin, eds., Springer-Verlag, New York, 1988, pp. 35–49.

[5] C. Atkinson, G. E. H. Reuter, and C. J. Ridler-Rowe, *Traveling wave solutions for some nonlinear diffusion equations*, SIAM J. Math. Anal., 12 (1981), pp. 880–892.

[6] F. V. Atkinson and L. A. Peletier, *Similarity profiles of flows through porous media*, Arch. Rational Mech. Anal., 42 (1971), pp. 369–379.

[7] G. I. Barenblatt, *On some unsteady motions of a liquid and a gas in a porous medium*, Prikl. Mat. Mekh., 16 (1952), pp. 67–78 (in Russian).

[8] P. Bénilan, *Evolution Equations and Accretive Operators*, Lecture notes taken by S. Lenhart, University of Kentucky, Lexington, KY, 1981.

[9] F. Bernis, J. Hulshof, and J. L. Vazquez, *A very singular solution for the dual porous medium equation and the asymptotic behaviour of general solutions*, J. Reine Angew. Math., 435 (1993), pp. 1–31.

[10] E. Comparini, R. Ricci, and J. L. Vazquez, *Asymptotic behavior of the solutions of a nonlinear Fokker-Planck equation with Dirichlet boundary conditions*, J. Math. Anal. Appl., 175 (1993), pp. 606–631.

[11] C. Cortazar and M. Elgueta, *Localization and boundedness of the solutions of the Neumann problem for a filtration equation*, Nonlinear Anal., 13 (1989), pp. 33–41.

[12] E. DiBenedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.

[13] E. DiBenedetto, *A boundary modulus of continuity for a class of singular parabolic equations*, J. Differential Equations, 63 (1986), pp. 418–447.

[14] B. H. Gilding, *Stabilization of flows through porous media*, SIAM J. Math. Anal., 10 (1979), pp. 237–246.

[15] B. H. Gilding and M. A. Herrero, *Localization and blow-up of thermal waves in nonlinear heat conduction with peaking*, Math. Ann., 282 (1988), pp. 223–242.

[16] B. H. Gilding, R. Natalini, and A. Tesei, *How parabolic free boundaries approximate hyperbolic fronts*, Trans. Amer. Math. Soc., 352 (2000), pp. 1797–1824.

[17] B. H. Gilding and L. A. Peletier, *On a class of similarity solutions of the porous media equation*, J. Math. Anal. Appl., 55 (1976), pp. 351–364.

[18] J. Goncerzewicz, *Properties of solutions to the initial-boundary value problem for a porous media-type equation*, Math. Methods Appl. Sci., 15 (1992), pp. 299–314.

[19] J. Goncerzewicz, *On initial-boundary value problems for a certain class of degenerate parabolic equations*, in preparation.

[20] J. Hulshof and J. L. Vazquez, *The dipole solution for the porous medium equation in several space dimensions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 20 (1993), pp. 193–217.

[21] A. S. Kalashnikov, *The occurrence of singularities in solutions of the non-steady seepage equation*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1967), pp. 269–275. Translation of Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 440–444.

[22] A. S. Kalashnikov, *On the differential properties of generalized solutions of equations of the nonsteady-state filtration type*, Moscow Univ. Math. Bull, 29(1) (1974), pp. 48–53. Translation of Vestnik Moskov. Univ. Ser. I Mat. Mekh., 29(1) (1974), pp. 62–68.

[23] A. S. Kalashnikov, *Some problems of the qualitative theory of non-linear degenerate second-order parabolic equations*, Russian Math. Surveys, 42(2) (1987), pp. 169–222. Translation of Uspekhi Mat. Nauk, 42(2) (1987), pp. 135–176.

[24] W. Mydlarczyk, *A singular initial value problem for second and third order differential equations*, Colloq. Math., 68 (1995), pp. 249–257.

[25] R. E. Pattle, *Diffusion from an instantaneous point source with a concentration-dependent coefficient*, Quart. J. Mech. Appl. Math., 12 (1959), pp. 407–409.

[26] L. A. Peletier, *Asymptotic behavior of solutions of the porous media equation*, SIAM J. Appl. Math., 21 (1971), pp. 542–551.

[27] L. A. Peletier, *The porous media equation*, in Applications of Nonlinear Analysis in the Physical Sciences, H. Amann, N. Bazley, and K. Kirchgässner, eds., Pitman Advanced Publishing Program, Boston, 1981, pp. 229–241.

[28] L. A. Peletier and A. Tesei, *Diffusion in inhomogeneous media: Localization and positivity*, Ann. Mat. Pura Appl. (4), 141 (1985), pp. 307–330.

[29] F. Quirós and J. L. Vázquez, *Asymptotic behaviour of the porous media equation in an exterior domain*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 28 (1999), pp. 183–227.

[30] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Localization of diffusion processes in media with constant properties*, Soviet Phys. Dokl., 24 (1979), pp. 543–545. Translation of Dokl. Akad. Nauk SSSR, 247 (1979), pp. 349–353.

[31] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-up in Quasilinear Parabolic Equations*, Walter de Gruyter, Berlin, 1995.

[32] A. E. Shishkov, *Boundary blow-up for energy solutions of general multi-dimensional parabolic equations*, Nelineinye Granichnye Zadachi, 8 (1998), pp. 229–237.

[33] A. E. Shishkov and A. G. Shchelkov, *Blow-up boundary regimes for general quasilinear parabolic equations in the multidimensional domains*, Sb. Math., 190 (1999), pp. 447–479.

[34] J. L. Vázquez, *Symétrisation pour $u_t = \Delta\varphi(u)$ et applications*, C. R. Acad. Sci. Paris Sér. I Math., 295 (1982), pp. 71–74.

[35] J. L. Vazquez, *Asymptotic behaviour and propagation properties of the one-dimensional flow of gas in a porous medium*, Trans. Amer. Math. Soc., 277 (1983), pp. 507–527.

[36] J. L. Vazquez, *An introduction to the mathematical theory of the porous medium equation*, in Shape Optimization and Free Boundaries, M. C. Delfour and G. Sabidussi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992, pp. 347–389.

[37] W. Walter, *Differential and Integral Inequalities*, Springer-Verlag, Berlin, 1970.

[38] Ya. B. Zel'dovich and A. S. Kompaneets, *On the theory of propagation of heat with the heat conductivity depending upon the temperature*, in Collection in Honour of the Seventieth Birthday of Academician A. F. Ioffe, Izdat. Akad. Nauk SSSR, Moscow, 1950, pp. 61–71 (in Russian).

SIAM J. Math. Anal.
Vol. 31, No. 4, pp. 894–908

BLOWUP FOR SYSTEMS OF CONSERVATION LAWS[*]

HELGE KRISTIAN JENSSEN[†]

Abstract. We give examples of finite time blowup in sup-norm and total variation for 3×3-systems of strictly hyperbolic conservation laws. The exact solutions are explicitly constructed. In the case of sup-norm blowup we also provide an example where all other p-norms, $1 \leq p < \infty$, remain uniformly bounded. Finally we consider appropriate rescalings for the different types of blowup.

Key words. systems of conservation laws, blowup, total variation, sup-norm, rescaling

AMS subject classifications. 35L65, 35B05

PII. S0036141099352339

1. Introduction.

We consider systems of conservation laws of the form

$$U_t + F(U)_x = 0 \tag{1.1}$$

with initial data

$$U(x, 0) = U_0(x), \tag{1.2}$$

where $U(x,t) = (u(x,t), v(x,t), w(x,t)) \in \mathbb{R}^3$ and $F : \mathbb{R}^3 \to \mathbb{R}^3$ is smooth and strictly hyperbolic; i.e., the Jacobian DF has real and distinct eigenvalues. We assume that each characteristic field is either genuinely nonlinear or linearly degenerate in the sense of Lax [21].

The existence of a weak entropy solution to the Cauchy problem for an $n \times n$-system of the form (1.1) has been established in two main cases. Either the total variation (T.V.) of the initial data is assumed to be sufficiently small, or one considers systems of two equations. In the seminal paper [15] Glimm introduced a functional consisting of a linear term giving the total variation of the solution and a quadratic term measuring the amount of waves generated by future collisions. For data close to a constant state and with small total variation the functional is decreasing in time, and a compactness argument yields a weak entropy solution to (1.1), (1.2). This solution is constructed by Glimm's scheme [15, 22] or by wave front-tracking [4, 5, 27].

Various extensions and refinements of the original result have been given. Young [33] proves a third-order estimate for wave interactions and uses this together with a reordering technique to obtain L^∞-stability for solutions constructed by Glimm's scheme. In [31] Temple and Young derive sufficient conditions for existence of solutions to 3×3-systems with a 2-Riemann invariant when the data has small amplitude but possibly large variation. The same class of systems is considered in [32] where existence of solutions up to any prescribed time is established for data with arbitrarily large total variation and correspondingly small sup-norm. In both this work and in the work by Cheverry, a new length scale for the Cauchy problem was introduced. Using this, Cheverry [10] has showed how to relax the restriction on the variation of the initial data for $n \times n$-systems where all the fields are genuinely nonlinear. Schochet [28] shows that for $n \times n$-systems with "almost planar interactions" the conditions on the

[*] Received by the editors February 22, 1999; accepted for publication July 27, 1999; published electronically April 4, 2000. This work was sponsored by NorFA.
http://www.siam.org/journals/sima/31-4/35233.html

[†] S.I.S.S.A., Via Beirut 4, Trieste 34014, Italy (jenssen@sissa.it).

initial data can be relaxed. Recently Bressan and Goatin [7] have considered Temple class systems where all fields are genuinely nonlinear. They prove the existence of an L^1-continuous semigroup for L^∞-data with possible infinite total variation.

For the case $n = 2$ stronger results have been obtained. Glimm and Lax [16] considered a large class of 2×2-systems and proved global existence of a weak solution under the much weaker assumption that the *oscillation* of the initial data is sufficiently small. Several works establish existence with large data for gas dynamics [11, 12, 23, 25, 26]. Similar results have been obtained by applying the theory of compensated compactness [13, 14]. Serre [30] has studied the case of 2×2 Temple class systems for which one has global existence for data with bounded variation. See [2] for L^1-continuous dependence in this case. Alber [1] has proved local existence for isentropic gas dynamics where the data have compact support and bounded variation. For an extension of this result to $n \times n$-systems, see [29]. A local uniqueness result in the case of small BV perturbations of (possibly large) Riemann data was established in [6]. For existence and continuous dependence for special systems with data in L^∞, see [3, 8].

Recent results on 3×3-systems by several authors [17, 18, 20, 24, 34] show that the restriction to 2×2-systems for these stronger results is essential. More specifically the authors consider the possibility of blowup in finite time of total variation or sup-norm. These works present special classes and explicit examples of systems for which different types of behavior can be found. In [18] Jeffrey gave an example of blowup in finite time of the sup-norm (and, hence, also of total variation) for a 3×3-system. The system is strictly hyperbolic and linearly degenerate in each characteristic family, and the solution is smooth. A weakness of this example is that the system is not in conservative form. Young [34] constructs exact solutions to 3×3-systems with periodic initial data. Depending on the choice of initial data and the interaction coefficients one gets different types of behavior. These include arbitrarily large magnifications of total variation and p-norms ($1 \leq p \leq \infty$) in finite time, decay rates like $1/(1 + t)$, exponential growth and decay, and time periodic solutions. The systems are linearly degenerate in each family (constant eigenvalues) so that the regularizing effect of genuine nonlinearity is absent and all nonlinear effects are due to the geometrical nonlinearities of the wave curves in state space. In [20] Joly, Metivier, and Rauch presented a class of systems which are genuinely nonlinear in all three fields. Using the theory of weakly nonlinear geometric optics, the authors show the existence of systems with periodic initial data where the variation grows arbitrarily large and the sup-norm is amplified by arbitrary large factors in finite time. A common feature of the works [20] and [34] is the use of initial data in BV_{loc} with sufficiently small amplitude such that the solution remains local in state space. In the former case, this guarantees that all Riemann problems can be solved uniquely, while in the latter it guarantees that the methods of weakly nonlinear geometric optics can be applied. For further work on the methods of nonlinear geometric optics applied to systems of conservation laws, see [9, 19, 24]. Recently Bressan and Shen [8] have given an example to the effect that the Cauchy problem is not well posed for 3×3-systems if one allows data with infinite total variation.

The works of Young and Joly, Metivier, and Rauch show that one cannot extend the Glimm–Lax theory to larger systems. In particular it is not possible to prove global existence for general 3×3-systems by deriving uniform BV-estimates. However, given these results it is still conceivable that one could obtain general global existence results for $n \times n$-systems, $n \geq 3$, even in the case of large data. We shall see that this is

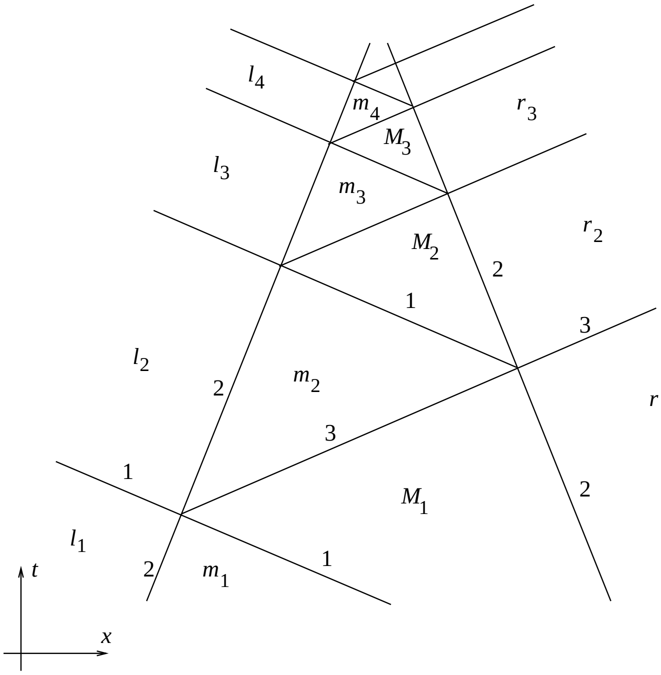


FIG. 1. *Interaction pattern.*

not possible. In what follows we present a class of 3×3-systems for which one can prescribe initial data such that the solution blows up in finite time. We will consider blowup in both sup-norm and total variation. That is, we give a class of examples for which there exists a time T, $0 < T < \infty$, such that

$$\lim_{t \to T^-} \|U(\cdot,t)\|_\infty = +\infty, \tag{1.3}$$

and we also give a class of examples for which there exists a time T, $0 < T < \infty$, such that

$$\lim_{t \to T^-} \text{T.V. } [U(\cdot,t)] = +\infty \quad \text{while} \quad \|U(\cdot,t)\|_\infty \quad \text{remains bounded.} \tag{1.4}$$

Both types of examples are constructed by considering a situation where two 2-shocks approach each other while 1- and 3-shocks are being reflected between the 2-shocks; see Figure 1. By carefully choosing the flux function and the initial data, we obtain the above behaviors. More precisely, we get examples where the waves are magnified at each interaction, which yields blowup in sup-norm, and where the solution is periodic in state space, yielding blowup in total variation. The examples differ from the examples given in [20], [34] both in the mechanism of blowup (i.e., the particular interaction pattern) and in the fact that one actually gets *infinite* sup-norm or total variation in *finite time*. To have blowup in sup-norm the 2-shocks must be sufficiently strong, and to have to blowup in total variation the strength of the 2-shocks must be chosen in a particular way to give periodicity in state space. However, the initial 1- and 3-waves can be arbitrarily weak. We also give an example where the sup-norm blows up while the p-norms of $U(\cdot,t)$ remain uniformly bounded as $t \to T^-$. These are, to the best of my knowledge, the first examples of this type.

The paper is organized as follows. In the next section we give the systems we will consider, we note their main properties, and we formulate the main result. In section 3 we consider Riemann problems and we derive a criteria for a Riemann problem to have a unique solution. Section 4 contains the proof of the main result. We also present rescalings which describe the asymptotics in the various cases of blowup. In the last section we collect some additional observations and comment on open problems.

2. Class of systems and statement of main result. We want to set up an interaction pattern like the one in Figure 1 where two 2-shocks approach each other while 1- and 3-waves (which will be contact discontinuities) are reflected back and forth between the 2-shocks. Note that this requires at least three equations; i.e., it is not possible to get an interaction pattern like this for 2×2-systems. We do this by constructing solutions to 3×3-systems of the form (1.1) where the flux function F has the form

$$F(U) \equiv F(u, v, w) = \begin{pmatrix} ua(v) + w \\ \Gamma(v) \\ u(\lambda_0^2 - a^2(v)) - wa(v) \end{pmatrix}. \tag{2.1}$$

Here $\lambda_0 > 0$ is a constant and $a(v)$ will be chosen later to obtain the types of behavior stated in Theorem 2.1. To simplify the analysis we assume that $\Gamma(v)$ has the following properties:

(i) $\Gamma(v)$ is strictly convex;

(ii) $-\lambda_0 < \gamma(v) \equiv \Gamma'(v) < \lambda_0$ for all $v \in \mathbb{R}$;

(iii) $\Gamma(0) = 0$ and $\Gamma(-v) = \Gamma(v)$ for all $v \in \mathbb{R}$.

It is readily checked that the eigenvalues of the Jacobian DF are

$$\lambda_1 = -\lambda_0, \qquad \lambda_2 = \gamma(v), \qquad \lambda_3 = +\lambda_0. \tag{2.2}$$

Thus (ii) guarantees that the system is strictly hyperbolic. Also, since the first and third eigenvalues are constants, the 1-waves to the left of the left 2-wave and the 3-waves to the right of the right 2-wave, respectively, do not interact (see Figure 1). The first and third eigenvectors are given by

$$r_1 = \begin{pmatrix} 1 \\ 0 \\ -(\lambda_0 + a(v)) \end{pmatrix}, \qquad r_3 = \begin{pmatrix} 1 \\ 0 \\ \lambda_0 - a(v) \end{pmatrix}. \tag{2.3}$$

Note that the second equation in the system is a decoupled scalar conservation law for v with a strictly convex flux. It follows that the second characteristic field is genuinely nonlinear. The first and third fields are linearly degenerate so that all 1-waves and 3-waves are contact discontinuities. It follows that shock and rarefaction curves coincide in the first and third families, and these are straight lines in planes with $v =$ constant.

Remark. This class of systems is a modification of the examples considered by Young [34]. The difference is that we have introduced nonlinearity in the second field. As in [34] we construct *exact* solutions.

We say that a solution of (1.1), (1.2) has an *interaction pattern* as in Figure 1 if the initial data consist of four constant states U_{l_1}, U_{m_1}, U_{M_1}, and U_{r_1} (ordered from left to right) and the Riemann problem

- (U_{l_1}, U_{m_1}) gives rise to a single 2-shock with positive speed;
- (U_{m_1}, U_{M_1}) gives rise to a single contact discontinuity of the first family;

- (U_{M_1}, U_{r_1}) gives rise to a single 2-shock with negative speed.

Note that the v-component does not change across 1- and 3-waves. Since the second equation is a scalar conservation law for v with a convex flux satisfying $\Gamma(0) = 0$, it follows that the solution has an interaction pattern as in Figure 1 if and only if $0 < v_{l_1} > v_{m_1} = v_{M_1} > v_{r_1} < 0$. We now state the main result of the paper.

THEOREM 2.1. *Let the flux F be given by* (2.1). *Then for a suitable choice of $a(v)$, $\Gamma(v)$, λ_0, and initial states U_{l_1}, U_{m_1}, U_{M_1}, and U_{r_1}, the solution of* (1.1), (1.2) *has an interaction pattern as in Figure* 1 *and satisfies one of the following relations.*

(a) *There exists a time T, $0 < T < \infty$, such that*

$$\lim_{t \to T^-} \|U(\cdot, t)\|_\infty = +\infty. \tag{2.4}$$

(b) *There exist a constant $C > 0$ and a time T, $0 < T < \infty$, such that*

$$\lim_{t \to T^-} \textit{T.V.}\ [U(\cdot, t)] = +\infty, \ \textit{while}\ \|U(\cdot, t)\|_\infty < C \ \textit{for all}\ t < T. \tag{2.5}$$

In the case of blowup in sup-norm one may choose the parameters so that all other p-norms of $U(\cdot, t)$, $1 \le p < \infty$, remain uniformly bounded as $t \to T^-$. Moreover, in all cases the system remains uniformly strictly hyperbolic in the sense that there is a $\delta > 0$ such that $-\lambda_0 + \delta < \gamma(v(x, t)) < \lambda_0 - \delta$ for all $(x, t) \in \mathbb{R} \times [0, T)$.

3. Riemann problems.

In this section we consider the Riemann problem for the system (1.1), i.e., the Cauchy problem when the data consist of two constant states,

$$U_0(x) = \begin{cases} U_l & \text{if } x < 0, \\ U_r & \text{if } x > 0. \end{cases} \tag{3.1}$$

We will consider only the case in which $v_l > v_r$ since this is all we need to construct solutions with the properties described in the theorem. The solution of the Riemann problem then consists of a contact discontinuity of the first family connecting U_l to some state U^-, followed by a 2-shock connecting U^- to some state U^+, followed by a contact discontinuity of the third family connecting U^+ to U_r.

We first parameterize the integral curves of the first and third family by s. These are straight lines and the parameterizations are readily obtained from (2.3). We let $D_j[s; (\bar{u}, \bar{v}, \bar{w})]$ denote the integral curve of the jth field, $j = 1, 3$, through the point $(\bar{u}, \bar{v}, \bar{w})$. Thus the first and third wave curves are given by

$$D_1[s; (\bar{u}, \bar{v}, \bar{w})] = \begin{pmatrix} s + \bar{u} \\ \bar{v} \\ -s(\lambda_0 + a(\bar{v})) + \bar{w} \end{pmatrix} \tag{3.2}$$

and

$$D_3[s; (\bar{u}, \bar{v}, \bar{w})] = \begin{pmatrix} s + \bar{u} \\ \bar{v} \\ s(\lambda_0 - a(\bar{v})) + \bar{w} \end{pmatrix}. \tag{3.3}$$

To find the expression for the 2-shock curve through $(\bar{u}, \bar{v}, \bar{w})$, we use the Rankine–Hugoniot condition. This states that if the solution contains a discontinuity with speed $\bar{\gamma}$, then

$$[F(U)] = \bar{\gamma}[U], \tag{3.4}$$

where $[\cdot]$ denotes the jump across the discontinuity. Let the left and right states be $(\bar{u},\bar{v},\bar{w})$ and (u,v,w), respectively. With the flux given by (2.1) the Rankine–Hugoniot condition takes the form

$$ua(v)+w-\bar{u}a(\bar{v})-\bar{w}=\bar{\gamma}(u-\bar{u}), \tag{3.5}$$

$$\Gamma(v)-\Gamma(\bar{v})=\bar{\gamma}(v-\bar{v}), \tag{3.6}$$

$$u(\lambda_0^2-a^2(v))-wa(v)-\bar{u}(\lambda_0^2-a^2(\bar{v}))+\bar{w}a(\bar{v})=\bar{\gamma}(w-\bar{w}). \tag{3.7}$$

These relations yield three curves. Along two of these v is constant and they coincide with D_1 and D_3. The third is the 2-shock curve for which we use v as a parameter. Given the point $(\bar{u},\bar{v},\bar{w})$, (3.6) gives the speed of the 2-shock,

$$\bar{\gamma}=\frac{\Gamma(v)-\Gamma(\bar{v})}{v-\bar{v}}. \tag{3.8}$$

Substituting this into (3.5) gives w expressed by u and v. Using this and (3.8) in (3.7) then yields u as a function of v. The expressions for u and w are given by

$$u=u(v;(\bar{u},\bar{v},\bar{w}))=\bar{u}+\frac{a(v)-a(\bar{v})}{\lambda_0^2-\bar{\gamma}^2}\Big(\bar{u}(a(\bar{v})-\bar{\gamma})+\bar{w}\Big) \tag{3.9}$$

and

$$w=w(v;(\bar{u},\bar{v},\bar{w}))=\bar{w}+(\bar{\gamma}-a(v))[u(v;(\bar{u},\bar{v},\bar{w}))-\bar{u}]-\bar{u}(a(v)-a(\bar{v})). \tag{3.10}$$

Note that these expressions are linear in $\bar{u}$ and $\bar{w}$.

We next use the solution of the Rankine–Hugoniot equations to derive a criteria to determine when the Riemann problem (U_l,U_r) has a unique solution. Let s_1 and s_3 denote the change in parameter across the 1-wave connecting U_l to U^- and the 3-wave connecting U^+ to U_r, respectively. Then

$$U^-=D_1[s_1;U_l], \tag{3.11}$$

$$U^+=\begin{pmatrix}u(v_r;U^-)\\ v_r\\ w(v_r;U^-)\end{pmatrix}, \tag{3.12}$$

and

$$U_r=D_3[s_3;U^+]. \tag{3.13}$$

This yields three equations, one for the speed of the 2-shock, given by

$$\bar{\gamma}=\frac{\Gamma(v_r)-\Gamma(v_l)}{v_r-v_l}, \tag{3.14}$$

while the other two are *linear* equations for the unknown strengths s_1 and s_3. These equations can be written in the form

$$s_1(\bar{\gamma}+\lambda_0)+s_3(\bar{\gamma}-\lambda_0)=A, \tag{3.15}$$

$$s_1[(\bar{\gamma}-\lambda_0)+(a(v_r)-a(v_l))]+s_3(\bar{\gamma}-\lambda_0)=B, \tag{3.16}$$

where A and B are functions of U_l, U_r. Thus the Riemann problem (U_l,U_r) has a unique solution if and only if

$$\lambda_0\neq\bar{\gamma},\ \text{ and } a(v_r)-a(v_l)-2\lambda_0\neq 0. \tag{3.17}$$

The first condition is always fulfilled by the mean value theorem and assumption (ii) on the flux Γ. In the proof of part (a) and (b) of Theorem 2.1 we will choose λ_0, v_r, v_l, and $a(v)$ such that the second condition is also satisfied at each interaction.

4. Proof of main result. We now construct examples with the properties stated in Theorem 2.1. Fix a $V > 0$ and let the v-component of the states to the left of the left 2-shock be V, let the v-component of the states between the two 2-shocks be 0, and let the v-component of the states to the right of the right 2-shock be $-V$. As noted above, this guarantees that the solution has an interaction pattern as in Figure 1. Also, since each state lies in one of the planes $v \equiv V$, $v \equiv 0$, or $v \equiv -V$, it is clear that the solution is uniformly strictly hyperbolic.

We next consider left interactions in which a 1-wave hits the left 2-shock from the right. Let the states l, m, M, l', and m' be as in Figure 2. Let the strength (i.e., the change in the parameter s) of the incoming 1-wave be S, while the transmitted 1-wave has strength T, and the reflected 3-wave has strength R. The given quantities are u_l, $v_l = v_{l'} = V$, w_l, $v_m = v_M = v_{m'} = 0$, and S, from which we want to compute the strengths T and R. Starting at l and going either via m or via l' and m' yield two expressions for the state M. This gives two linear equations for the strengths T an R. We have

$$M = D_1[S; m] = D_1[S; (u(0; l), 0, w(0; l))] \tag{4.1}$$

and

$$M = D_3[R; m'] = D_3[R; (u(0; l'), 0, w(0; l'))], \tag{4.2}$$

where $l' = D_1[T; l]$. We denote the speed of the left 2-shock by $\bar{\gamma}$, i.e.,

$$\bar{\gamma} = \frac{\Gamma(0) - \Gamma(V)}{-V} = \frac{\Gamma(V)}{V}. \tag{4.3}$$

We solve (4.1), (4.2) for T and R by using the expressions for the wave curves from above. A straightforward calculation yields

$$T = \alpha S, \qquad R = \beta S, \tag{4.4}$$

where the magnification coefficients α and β are given by

$$\alpha = \frac{2\lambda_0}{2\lambda_0 + a(V) - a(0)}, \qquad \beta = \frac{\lambda_0 + \bar{\gamma}}{\lambda_0 - \bar{\gamma}} \left(\frac{a(0) - a(V)}{2\lambda_0 + a(V) - a(0)} \right). \tag{4.5}$$

Note that these coefficients depend only on λ_0, a, and V. Next consider the situation where a 3-wave hits the right 2-wave. Let the states l, M, M', r', and r be as in Figure 3, and let the strength of the incoming 3-wave, the reflected 1-wave, and the transmitted 3-wave be S, R, and T, respectively. The given quantities are now u_l, $v_l = v_M = v_{M'} = 0$, w_l, $v_r = v_{r'} = -V$, and S, from which we want to compute the strengths T and R. Starting at l and going either via M or via M' and r' yield two expressions for the state r. This gives two linear equations for the strengths T and R. We have

$$r = (u(-V; M), -V, w(-V; M)),$$

where $M = D_3[S; l]$. Also

$$r = D_3[T; r'] = D_3[T; (u(-V; M'), -V, w(-V; M'))],$$

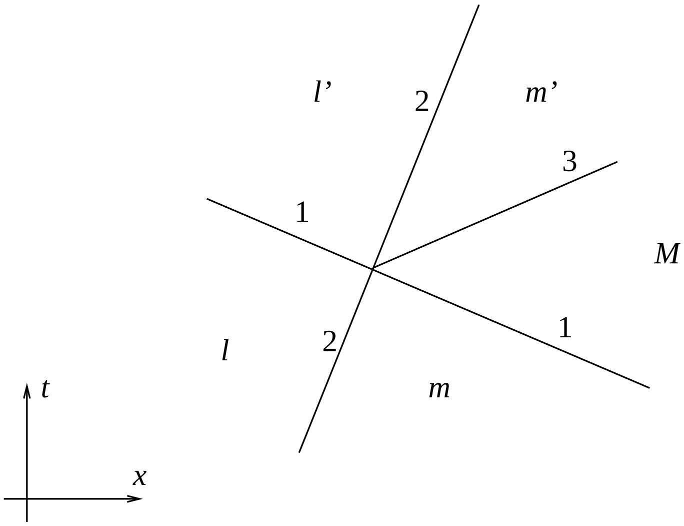


Fig. 2. *Left interaction.*

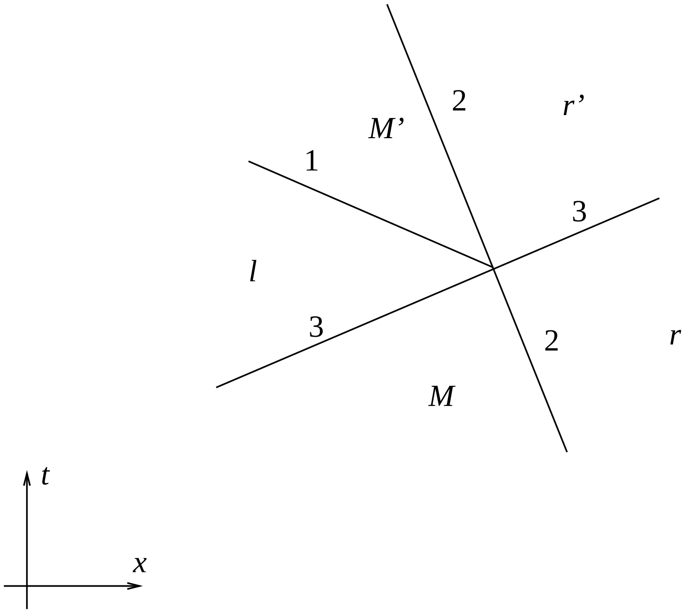


Fig. 3. *Right interaction.*

where $M' = D_1[R; l]$. We denote the speed of the right 2-shock by $\tilde{\gamma}$. By the properties of Γ we have

$$\tilde{\gamma} = \frac{\Gamma(-V) - \Gamma(0)}{-V} = -\frac{\Gamma(V)}{V} = -\bar{\gamma}. \tag{4.6}$$

Using the expressions for the wave curves we have

$$T = \delta S, \qquad R = \varepsilon S, \tag{4.7}$$

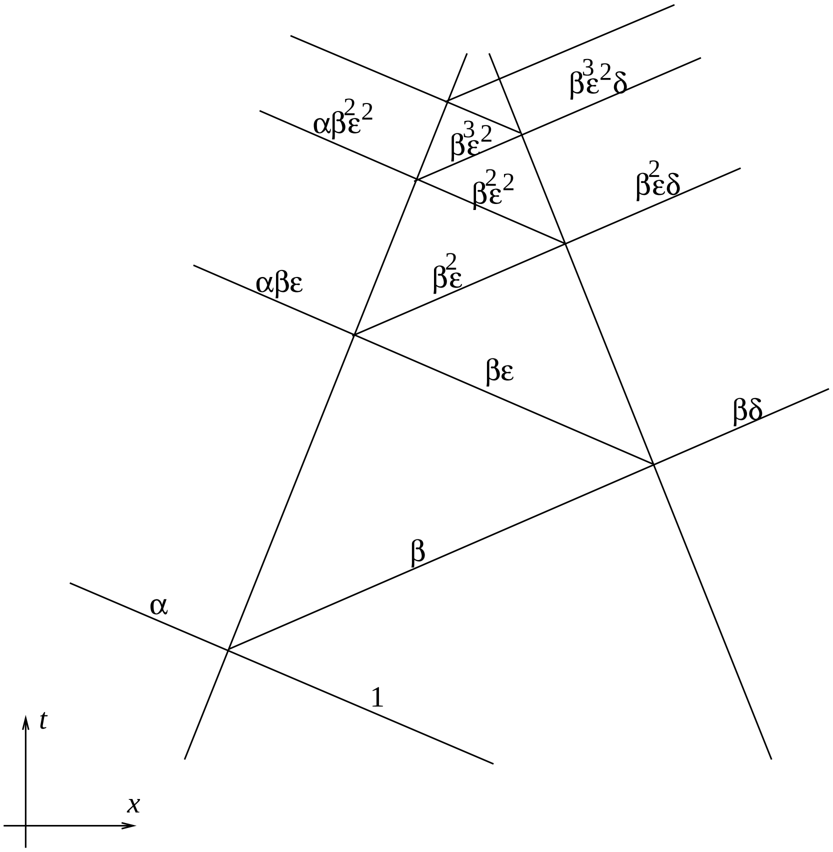


FIG. 4. *Wave strengths when first incoming* 1*-wave has strength* 1.

where the magnification coefficients δ and ε are given by

$$\delta = \frac{2\lambda_0}{2\lambda_0 + a(0) - a(-V)}, \qquad \varepsilon = \frac{\lambda_0 + \bar{\gamma}}{\lambda_0 - \bar{\gamma}} \left(\frac{a(-V) - a(0)}{2\lambda_0 + a(0) - a(-V)} \right). \tag{4.8}$$

As for the left interaction, the coefficients depend only on λ_0, a, and V.

Figure 4 shows the strengths of the various waves assuming that the first incoming 1-wave has strength 1.

We are now ready to choose a, V, and λ_0 so that the solution has the behavior stated in Theorem 2.1. We choose $\lambda_0 = 1$, and we let

$$\Gamma(v) = \frac{2}{\pi} \int_0^v \arctan(\xi)\, d\xi, \tag{4.9}$$

for which the properties (i)–(iii) are satisfied.

To prove part (a) we assume that $a(v) = v$. With this choice we have

$$\alpha = \delta \text{ and } \beta = \varepsilon.$$

Also, since $V > 0$, the criteria (3.17) is fulfilled, so that every Riemann problem occurring can be solved uniquely. We now refer to Figure 4 and observe that if

$|\beta| = |\varepsilon| > 1$, then the strengths of the 1- and 3-waves grow exponentially as a function of the number of interactions. Since there is an infinite number of interactions in finite time, it follows that the sup-norm tends to infinity in finite time provided $|\beta| > 1$. We have

$$|\beta| = \left(\frac{1+\bar{\gamma}}{1-\bar{\gamma}}\right)\left(\frac{V}{2+V}\right).$$

Since

$$\lim_{V\to\infty} \bar{\gamma} = 1,$$

it follows that $|\beta| > 1$ for V large enough. This completes the proof of part (a) of Theorem 2.1. An alternative is to choose V so that $\beta = \varepsilon = -1$. In this case the strengths of the transmitted 1- and 3-waves are constant and the sup-norm increases linearly as a function of the number of interactions taken place.

To prove part (b) we assume that $a(v) = v^2$. The criteria (3.17) is satisfied for the interactions along the left 2-shock, while it is satisfied for the interactions along the right 2-shock if and only if $|V| \neq \sqrt{2}$. Referring to Figures 1 and 4, we observe that

$$\begin{gathered}
l_1 = l_3 = l_5 = \ \cdots\ ,\ l_2 = l_4 = l_6 = \ \cdots\ ,\\
m_1 = m_3 = m_5 = \ \cdots\ ,\ m_2 = m_4 = m_6 = \ \cdots\ ,\\
M_1 = M_3 = M_5 = \ \cdots\ ,\ M_2 = M_4 = M_6 = \ \cdots\ ,\\
r_1 = r_3 = r_5 = \ \cdots\ ,\ r_2 = r_4 = r_6 = \ \cdots
\end{gathered}$$

if and only if the magnification factors β and ε satisfy

$$\beta\varepsilon = -1.$$

Thus, if we can choose $V \neq \sqrt{2}$ such that $\beta\varepsilon = -1$, then the solution is periodic in state space. With $a(v) = v^2$ we have

$$\beta\varepsilon = -\left(\frac{1+\bar{\gamma}}{1-\bar{\gamma}}\right)^2 \frac{V^4}{4-V^4}.$$

With $\Gamma(v)$ as above it is easily established that the equation $\beta\varepsilon = -1$ has a unique positive solution $V = V^*$. Also, $V^* \neq \sqrt{2}$ so that the criteria (3.17) is satisfied. We thus have a solution which is periodic in state space. Since there is an infinite number of interactions in finite time, it follows that the total variation tends to infinity in finite time while the sup-norm remains bounded. This completes the proof of part (b) of Theorem 2.1.

4.1. $\mathbf{L^p}$-norms. Having established the existence of solutions which blow up in either sup-norm or total variation, it is interesting to see whether one can have blowup in sup-norm while all other p-norms ($1 \leq p < \infty$) of $U(\cdot, t)$ remain bounded. Of course, as $U(\cdot, t)$ takes constant nonzero values outside large enough compact intervals, this refers to the p-norms computed over some compact interval. We shall see that this is indeed the case where the sup-norm increases as slowly as possible. This corresponds to the case noted above where the sup-norm increases linearly as a function of the number of collisions, i.e., when $\beta = \varepsilon = -1$. We give initial data such that the solution has an interaction pattern as in Figure 5. That is, at time $t = 0$ a 2-shock and a 3-wave (of strength 1) start at $x = -L$ and another 2-shock starts at $x = +L$. Denote the speeds of the 2-shocks by $\pm\bar{\gamma}$, and let x_n, t_n be the coordinates of the nth interaction. We have

$$x_n = (-1)^{n+1} b^n L, \qquad t_n = \frac{L}{\bar{\gamma}}(1 - b^n), \tag{4.10}$$

where

$$b = \frac{\lambda_0 - \bar{\gamma}}{\lambda_0 + \bar{\gamma}}. \tag{4.11}$$

Referring to Figure 5 and using the expressions for the 1- and 3-wave curves, one checks that the states in this case are given as follows:

$$l_{n+1} = l_1 - \alpha n r_1(+V), \tag{4.12}$$

$$m_{n+1} = m_1 + n[r_3(0) - r_1(0)], \tag{4.13}$$

$$M_{n+1} = M_1 + n[r_3(0) - r_1(0)], \tag{4.14}$$

$$r_{n+1} = r_1 + \alpha n r_3(-V). \tag{4.15}$$

Now let t be a time between t_{2n} and t_{2n+1}. It is readily checked that the part of $\|U(\cdot,t)\|_p^p$ corresponding to the part of the solution between the two 2-shocks is bounded by a term of the form $Cb^n n^p$, while the part of $\|U(\cdot,t)\|_p^p$ corresponding to the solution to the left and right of the two 2-shocks are bounded by sums of the form

$$C \sum_{k=1}^{n} b^{2k} k^p. \tag{4.16}$$

Since $0 < b < 1$, this shows that the p-norms are indeed bounded for all values of $p \in [1, \infty)$. This completes the proof of the theorem.

5. Rescalings and time-periodic solutions.

A standard technique for studying blowup phenomena is to introduce rescaled coordinates. One seeks rescalings of both the independent and dependent variables so that the rescaled solution is nontrivial and more easily described.

Consider the type of blowup described by part (b) of Theorem 2.1. Since the solution in this case is periodic in state space, a natural question is whether one can find a rescaling of the independent variables which yields a *time-periodic* solution to a corresponding 3×3-system of hyperbolic equations. We will briefly describe a suitable rescaling which describes the blowup of case (b) of Theorem 2.1. We will find that the rescaled solution is periodic for large enough times on every compact interval.

Again, we give initial data such that the solution has an interaction pattern as in Figure 5. We denote the blowup time by T, i.e., $T = L/\bar{\gamma}$. Denoting the rescaled time by $\tau = \tau(t)$, we want $\tau(t_{n+1}) - \tau(t_n)$ to be constant and equal to twice the period. The simplest way of obtaining this is to define τ by

$$\tau = -\ln(T - t). \tag{5.1}$$

To have a periodic solution we must rescale the space variable such that the curves corresponding to the 2-shocks are vertical straight lines. Denoting the new space variable by η, we rescale the x-variable as follows:

$$\eta = \frac{x}{T - t}. \tag{5.2}$$

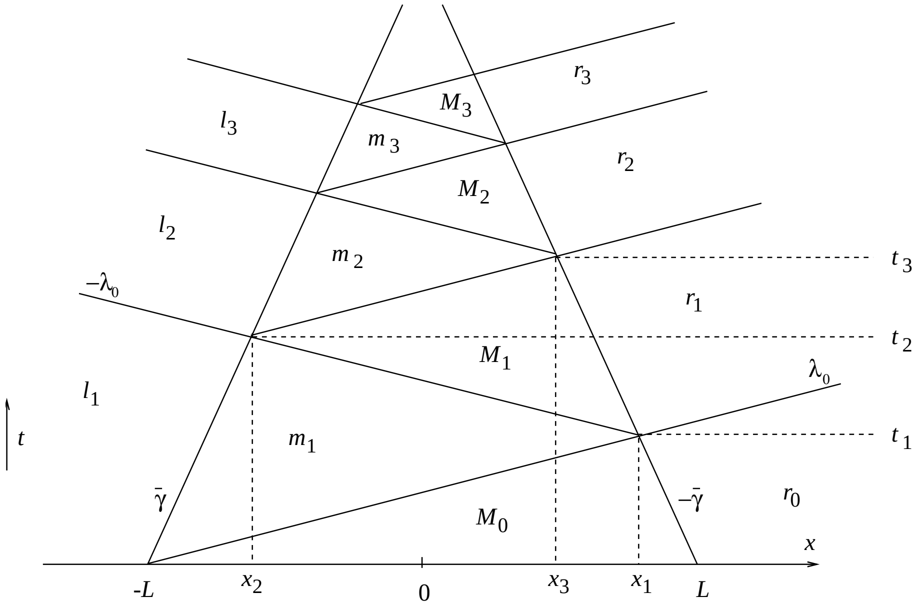


FIG. 5. *Before scaling.*

The straight lines of the 2-shocks are then mapped to the two lines $\eta \equiv \pm L/T = \pm\bar{\gamma}$, while the straight lines of the contact discontinuities are mapped to τ-translates of exponential curves of the form

$$\eta(\tau) = \pm[T(\lambda_0 - \bar{\gamma})e^{\tau} - \lambda_0].$$

The corresponding solution is then τ-periodic with period $-2\ln b$; see Figure 6. Thus the solution of (1.1) for which we have blowup in total variation may alternatively be described as a solution of the rescaled system

$$U_\tau + \eta U_\eta + F(U)_\eta = 0, \tag{5.3}$$

which is such that given any compact interval there is a time after which the solution is time-periodic on this interval.

For the case of blowup in sup-norm one has a similar result. We describe this without going into details. The scaling of the independent variables is again given by (5.1) and (5.2), while the scaling of the dependent variable is different in the two cases $|\beta| > 1$ and $\beta = -1$. The new dependent variables should be

$$\tilde{U} = \frac{U}{(T-t)^\rho}, \qquad \tilde{U} = -\frac{U}{\ln(T-t)},$$

respectively. Here $\rho = \ln\beta^2/\ln b$. The rescaled solution will tend to constant values on the η–intervals $(-\infty, -\bar{\gamma})$ and $(\bar{\gamma}, \infty)$, while it is time-periodic with period $-2\ln b$ on the middle interval $[-\bar{\gamma}, \bar{\gamma}]$.

6. Additional observations. Consider the system (1.1) with flux (2.1) obtained by replacing $\Gamma(v)$ with

$$\Gamma(v,k) = \frac{2}{\pi}\int_0^v \arctan(k\xi)\,d\xi. \tag{6.1}$$

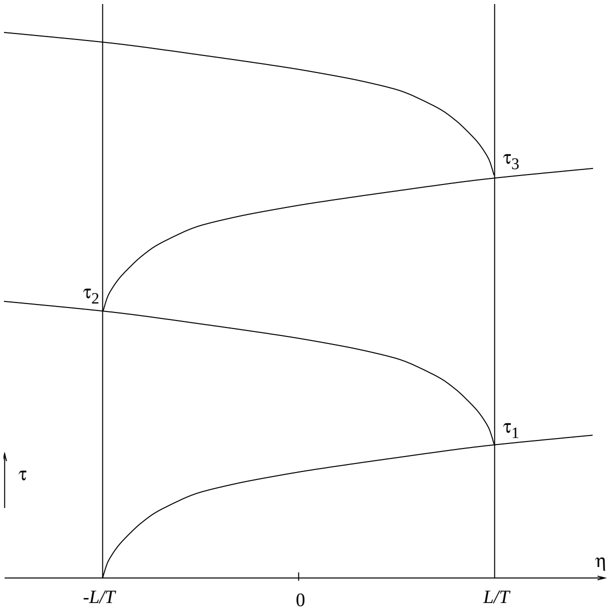


Fig. 6. *After scaling.*

By choosing large values for k it is easy to show that for any $\delta > 0$ one can find a system with initial data U_0 for which either

(i) $\|U_0\|_\infty < \delta$ and the solution of (1.1), (1.2) satisfies the conclusion of part (a) of Theorem 2.1

or

(ii) T.V. $[U_0] < \delta$ and the solution of (1.1), (1.2) satisfies the conclusion of part (b) of Theorem 2.1.

However, if one has an interaction pattern as in Figure 1, then the total variation of the initial data is bounded by $6\|U_0\|_\infty$. It follows by Glimm's result that it is impossible to find a *fixed* system with an interaction pattern as in Figure 1 and with the property that given any $\delta > 0$, there is a U_0 with $\|U_0\|_\infty < \delta$ and such that either of the behaviors in Theorem 2.1 occur.

We observe that the presence of infinitely many interactions in finite time does not necessarily imply that the solution ceases to exist. For example, in the case where $a(v) = v$, if we choose V so that $|\beta| < 1$, then the states to the left of the left 2-shock and the states to the right of the right 2-shock will converge to some states l_∞ and r_∞, respectively. These states then define a new Riemann problem at time $t = T$ at the point where the two 2-shocks meet. Solving this yields a 1-wave with speed $-\lambda_0$, a 3-wave with speed $+\lambda_0$, and a 2-shock with an intermediate speed. All other 1- and 3-waves from earlier interactions are prolonged and do not interact.

We observe that the type of interaction pattern as in Figure 1 is exactly what goes wrong with front-tracking for systems if one does not include a simplified Riemann solver for weak interactions. If one were to try to solve each Riemann problem in an exact manner (by solving for shocks exactly and by approximating rarefaction waves with many small shocks), then the examples above show that one may end up with infinitely many fronts in finite time. See [3, 5] for the definition of simplified Riemann solvers.

The systems considered above are quite artificial. First of all it would be interesting to obtain similar results for systems where all fields are genuinely nonlinear. In this case there are additional problems due to the possible interaction of transmitted 1- or 3-waves. The waves created in these interactions could interact with the 2-waves before infinitely many fronts have been created, and the analysis would be more complicated. Another basic problem is to understand the role played by entropies. Apart from the fact that the existence of a (strictly) convex entropy precludes blowup in L^2, very little seems to be known.

Also note that the systems we consider are in some sense opposite to the most interesting physical example of gas dynamics. For the Euler equations, the first and third fields are genuinely nonlinear while the second field is linearly degenerate. In this case, for initial data with large total variation and correspondingly small sup-norm, the results of Temple and Young [31, 32] show that one cannot have more than exponential growth in total variation. The problem of whether blowup in sup-norm is possible for gas dynamics when the data have large sup-norm remains open.

Acknowledgments. I thank A. Bressan for suggesting this problem. I am indebted to B. Piccoli, C. Dafermos, H. Holden, C. Cheverry, and G.-Q. Chen for enlightening discussions. This work was completed at the Mittag-Leffler Institute, and I thank director K. O. Widman and his staff for providing excellent working conditions.

REFERENCES

[1] H.-D. Alber, *Local existence of weak solutions to the quasilinear wave equation for large initial values*, Math. Z., 190 (1985), pp. 249–276.

[2] P. Baiti and A. Bressan, *The semigroup for Temple class system with large data*, Differential Integral Equations, 10 (1997), pp. 401–418.

[3] P. Baiti and H. K. Jenssen, *Well-posedness for a class of* 2×2 *conservation laws with* L^∞ *data*, J. Differential Equations, 140 (1997), pp. 161–185.

[4] P. Baiti and H. K. Jenssen, *On the front tracking algorithm*, J. Math. Anal. Appl., 217 (1998), pp. 395–404.

[5] A. Bressan, *Global solutions to systems of conservation laws by wave-front tracking*, J. Math. Anal. Appl., 170 (1992), pp. 414–432.

[6] A. Bressan and R. M. Colombo, *Unique solutions of* 2×2 *conservation laws with large data*, Indiana Univ. Math. J., 44 (1995), pp. 677–725.

[7] A. Bressan and P. Goatin, *Stability of* L^∞ *solutions of Temple class systems*, Differential Integral Equations, to appear; also available online from http://www.math.ntnu.no/conservation/.

[8] A. Bressan and W. Shen, *Uniqueness for discontinuous O.D.E. and conservation laws*, Nonlinear Anal., 34 (1998), pp. 637–652.

[9] C. Cheverry, *Justification de l'optique géométrique non linéaire pour un système de lois de conservation*, Duke Math. J., 87 (1997), pp. 213–263.

[10] C. Cheverry, *Système de lois de conservation et stabilité BV*, Mém. Soc. Math. France (N.S.), 75 (1998), pp. 1–106.

[11] R. J. DiPerna, *Global solutions to a class of nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 26 (1973), pp. 1–28.

[12] R. J. DiPerna, *Existence in the large for quasilinear hyperbolic conservation laws*, Arch. Rational Mech. Anal., 52 (1973), pp. 244–257.

[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] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure. Appl. Math., 18 (1965), pp. 697–715.

[16] J. Glimm and P. D. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc. 101, AMS, Providence, RI, 1970.

[17] J. K. Hunter, *Strongly nonlinear hyperbolic waves*, in Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications, J. Ballman and R. Jeltsch, eds., Notes Numer. Fluid Mech. 24, Vieweg, Braunschweig, 1989, pp. 257–268.

[18] A. Jeffrey, *Breakdown of the solution to a completely exceptional system of hyperbolic equations*, J. Math. Anal. Appl., 45 (1974), pp. 375–381.

[19] J. L. Joly, G. Metivier, and J. Rauch, *Resonant one-dimensional nonlinear geometric optics*, J. Funct. Anal., 114 (1993), pp. 106–231.

[20] J. L. Joly, G. Metivier, and J. Rauch, *A nonlinear instability for* 3×3 *systems of conservation laws*, Comm. Math. Phys., 162 (1994), pp. 47–59.

[21] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[22] T. P. Liu, *The deterministic version of the Glimm scheme*, Comm. Math. Phys., 57 (1977), pp. 135–148.

[23] T. P. Liu, *Solutions in the large for the equations of nonisentropic gas dynamics*, Indiana Univ. Math. J., 26 (1977), pp. 147–177.

[24] A. Majda and R. Rosales, *Resonantly interacting weakly nonlinear hyperbolic waves.* I. *A single space variable*, Stud. Appl. Math., 71 (1984), pp. 149–179.

[25] T. Nishida, *Global solution for an initial boundary value problem of a quasilinear hyperbolic system*, Proc. Japan Acad., 44 (1968), pp. 642–646.

[26] T. Nishida and J. A. Smoller, *Solutions in the large for some nonlinear hyperbolic conservation laws*, Comm. Pure Appl. Math., 26 (1973), pp. 183–200.

[27] N. H. Risebro, *A front-tracking alternative to the random choice method*, Proc. Amer. Math. Soc., 117 (1993), pp. 1125–1139.

[28] S. Schochet, *Glimm's scheme for systems with almost-planar interactions*, Comm. Partial Differential Equations, 16 (1991), pp. 1423–1440.

[29] S. Schochet, *Sufficient conditions for local existence via Glimm's scheme for large BV data*, J. Differential Equations, 89 (1991), pp. 317–354.

[30] D. Serre, *Solutions à variations bornées pour certains systèmes hyperboliques de lois de conservation*, J. Differential Equations, 68 (1987), pp. 137–168.

[31] B. Temple and R. Young, *Solutions to the Euler equations with large data*, in Hyperbolic Problems: Theory, Numerics Applications, World Scientific, River Edge, NJ, 1996, pp. 258–267.

[32] B. Temple and R. Young, *The large time stability of sound waves*, Comm. Math. Phys., 179 (1996), pp. 417–466.

[33] R. Young, *Sup-norm stability for Glimm's scheme*, Comm. Pure Appl. Math., 46 (1993), pp. 903–948.

[34] R. Young, *Exact solutions to degenerate conservation laws*, SIAM. J. Math. Anal., 30 (1999), pp. 537–558.

SIAM J. Math. Anal.
Vol. 31, No. 4, pp. 909–924

ON THE MULTIPLICITY OF SOLUTIONS OF TWO NONLOCAL VARIATIONAL PROBLEMS[*]

XIAOFENG REN[†] AND JUNCHENG WEI[‡]

Abstract. We study two nonlocal variational problems in this paper. One models microphase separation of diblock copolymers and the other models solid-solid phase transformations that lead to fine structures. We study a parameter range where the problems can be approximated by their asymptotic limits. We find all the local minimum solutions of the limiting problems. Because these local minima are isolated, and hence stable under perturbation, near them there exist local minimum solutions of the original problems.

Key words. nonlocal variational problems, Γ-convergence, isolated local minima, *BV* functions

AMS subject classifications. 49K15, 49K20, 34D15

PII. S0036141098348176

1. Introduction.

Two nonlocal variational problems are studied in this paper. The first one is

$$\mathcal{I}_\epsilon(u) = \begin{cases} \displaystyle\int_\Omega \left\{ \frac{\epsilon}{2}|\nabla u|^2 + \frac{1}{\epsilon}W(u) + \frac{1}{2}|(-\gamma^2\Delta)^{-1/2}(u-m)|^2 \right\} dx, \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } u \in \mathcal{A}_m \cap W^{1,2}(\Omega), \\ \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{if } u \in \mathcal{A}_m \backslash W^{1,2}(\Omega). \end{cases} \tag{1.1}$$

Ω is a bounded and smooth domain in R^d. ϵ and γ are both positive numbers. ϵ is small and γ is fixed.

W is a balanced double-well function with two global minima at -1 and 1, i.e., $W(t) \geq 0$ and $W(t) = 0$ if and only if $t = -1$ or 1. We also assume that W is continuous and there exist $k \geq 2$, $K_1 > 0$, $K_2 > 0$, and $\bar{t} > 1$ such that for all t, $|t| > \bar{t}$,

$$K_1|t|^k \leq W(t) \leq K_2|t|^k. \tag{1.2}$$

$-\gamma^2\Delta$ is the Laplacian operator, multiplied by the constant $-\gamma^2$, with the Neumann boundary condition. The outward normal vector field of $\partial\Omega$ is denoted by n. It is known that the operator

$$-\gamma^2\Delta : \left\{ u \in W^{2,2}(\Omega) : \int_\Omega u = 0, \ \frac{\partial u}{\partial n}|_{\partial\Omega} = 0 \right\} \to \left\{ u \in L^2(\Omega) : \int_\Omega u = 0 \right\}$$

is an isomorphism. The inverse of $-\gamma^2\Delta$ is self-adjoint and positive. We denote its positive square root by $(-\gamma^2\Delta)^{-1/2}$. This is a nonlocal operator.

The admissible set of $\mathcal{I}_\epsilon$ is

$$\mathcal{A}_m = \left\{ u \in L^2(\Omega) : \frac{1}{|\Omega|}\int_\Omega u\, dx = m \right\} \tag{1.3}$$

[*] Received by the editors December 2, 1998; accepted for publication (in revised form) July 28, 1999; published electronically April 4, 2000.
http://www.siam.org/journals/sima/31-4/34817.html

[†] Department of Mathematics and Statistics, Utah State University, Logan, UT 84322 (ren@math.usu.edu). This author was supported in part by NSF grant DMS-9703727.

[‡] Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). This author was supported in part by an Earmarked Grant of RGC of Hong Kong.

with the restriction $m \in (-1,1)$. For a measurable set Ω in R^d, $|\Omega|$ denotes its Lebesgue measure. In $\mathcal{A}_m$ we use the metric so that the distance between u and v is $\|u-v\|_2$, the $L^2(\Omega)$-norm of $u-v$. The choice of $L^2(\Omega)$ (as opposed to the choice of $L^1(\Omega)$ in the literature of Γ-convergence) is natural because of the nonlocal part, $(-\gamma^2\Delta)^{-1/2}$, of the functional.

In the study of diblock copolymers, a model was introduced in Ohta and Kawasaki [11] and Bahiana and Oono [1]. It asserts that the free energy of a diblock copolymer takes the form

$$\begin{aligned}&\mathcal{F}_{\epsilon,\sigma}(u) = \int_\Omega \left\{ \frac{\epsilon^2}{2}|\nabla u|^2 + W(u) + \frac{\sigma}{2}|(-\Delta)^{-1/2}(u-m)|^2 \right\} dx,\\ &\frac{1}{|\Omega|}\int_\Omega u\, dx = m.\end{aligned}$$

In a diblock copolymer, a linear-chain molecule consists of two subchains grafted covalently to each other. The subchains are made of two different monomer units, represented by $u=-1$ and $u=1$, respectively. The different subchains tend to segregate below some critical temperature, but as they are chemically bonded, only local microphase separation occurs. The connectivity of the two monomer units leads to the long range interaction term $\frac{\sigma}{2}|(-\Delta)^{-1/2}(u-m)|^2$ in the free energy. The parameter σ is proportional to the inverse of the square root of the total chain length of the copolymer. $\frac{\epsilon^2}{2}|\nabla u|^2$ represents the interfacial energy density at bonding points. The parameter ϵ is proportional to the thickness of interfaces between the two monomers. The double-well potential W prefers segregated monomers to a mixture. m stands for the mass ratio of the two monomer units. When this free energy is being minimized, the first term prefers large blocks of monomers, therefore reducing the combined size of interfaces between the two monomers. The third term, on the other hand, likes rapid oscillation between the two monomers. These two tendencies are competing. The process of reaching a stable configuration is known as microseparation. The results of our paper show that in a parameter range, namely $0 < \epsilon \approx \sigma \ll 1$, the monomer components in the copolymer develop blocks of a finite scale. For more references on the mathematical aspects of diblock copolymers we refer the reader to Nishiura and Ohnishi [10], where a different parameter range, $0 < \epsilon \ll \sigma \ll 1$, is studied. The functional studied in Müller [9] can also be written in the form of this model with $m=0$. There the parameter range is $0 < \epsilon \ll \sigma \approx 1$.

The second problem is

$$\mathcal{J}_\epsilon(u) = \begin{cases} \displaystyle\int_\Omega \left\{ \frac{\epsilon}{2}|\nabla u|^2 + \frac{1}{\epsilon}W(u) - \frac{1}{2}u^2 + \frac{1}{2}|(-\gamma^2\Delta+1)^{-1/2}u|^2 \right\} dx, & \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } u \in \mathcal{A}_m \cap W^{1,2}(\Omega), & \\ \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{if } u \in \mathcal{A}_m \backslash W^{1,2}(\Omega) & \end{cases} \tag{1.4}$$

in the same admissible set $\mathcal{A}_m$, (1.3). W satisfies the same conditions as in $\mathcal{I}_\epsilon$. The operator

$$-\gamma^2\Delta + 1 : \ \left\{ u \in W^{2,2}(\Omega) : \frac{\partial u}{\partial n}|_{\partial\Omega} = 0 \right\} \to L^2(\Omega)$$

is an isomorphism. Again we denote the positive square root of the inverse of $-\gamma^2\Delta+1$ by $(-\gamma^2\Delta+1)^{-1/2}$.

In studying solid-solid phase transformations, Ren and Truskinovsky proposed in [12] a model of 1-dimensional elastic bars that develop a mixture of two phase variants. Let u be the strain field of a deformed elastic bar. The stored energy is

$$\mathcal{F}_\epsilon(u) = \int_0^1 \left\{ \frac{\epsilon^2}{2}|\nabla u|^2 + W(u) - \frac{\epsilon}{2}u^2 + \frac{\epsilon}{2}|(-\gamma^2\Delta + 1)^{-1/2}u|^2 \right\} dx,$$

$$\int_0^1 u\,dx = m,$$

where m is the total displacement of the bar. $W(u)$ is the local part of the energy density. It is assumed to be nonconvex. It prefers the two phase variants, $u = -1$ and $u = 1$. $\frac{\epsilon^2}{2}|\nabla u|^2$ is the short range self-interaction of the strain field and $-\frac{\epsilon}{2}u^2 + \frac{\epsilon}{2}[(-\gamma^2\Delta + 1)^{-1/2}u]^2$ is the long range self-interaction. Similar to the copolymer model, these two competing factors lead to a mixture of the two phase variants. The result of this paper again shows the characteristic scale of each phase in the mixture.

When we study $\mathcal{I}_\epsilon$ and $\mathcal{J}_\epsilon$, we will show that as ϵ tends to 0, $\mathcal{I}_\epsilon$ and $\mathcal{J}_\epsilon$ both converge to their limiting problems, defined in (2.1) and (4.1). The convergence falls in the general theory of Γ-limits.

Then we will study the 1-dimensional cases, $\Omega = (0,1)$, of $\mathcal{I}_\epsilon$ and $\mathcal{J}_\epsilon$. We find all local minima of the limiting problems. It turns out that these local minima of the limiting problems are isolated, and near them there are local minima of $\mathcal{I}_\epsilon$ and $\mathcal{J}_\epsilon$ if ϵ is sufficiently small.

For each positive integer ν, we set

$$x_1 = \frac{1-m}{2\nu},\ x_2 = x_1 + \frac{1+m}{\nu},\ x_3 = x_2 + \frac{1-m}{\nu}, \dots,\ x_\nu = x_{\nu-1} + \frac{1+(-1)^\nu m}{\nu}.$$

We also set $x_0 = 0$ and $x_{\nu+1} = 1$. We define a step function $U_{\nu,1} \in \mathcal{A}_m$ so that

$$U_{\nu,1}(x) = (-1)^i \text{ if } x \in (x_{i-1}, x_i),\ i = 1, 2, \dots, \nu+1. \tag{1.5}$$

In a similar way for each positive integer ν, we set

$$z_1 = \frac{1+m}{2\nu},\ z_2 = z_1 + \frac{1-m}{\nu},\ z_3 = z_2 + \frac{1+m}{\nu}, \dots,\ z_\nu = z_{\nu-1} + \frac{1-(-1)^\nu m}{\nu}.$$

We also set $z_0 = 0$ and $z_{\nu+1} = 1$. We define $U_{\nu,2} \in \mathcal{A}_m$ so that

$$U_{\nu,2}(x) = (-1)^{(i-1)} \text{ if } x \in (z_{i-1}, z_i),\ i = 1, 2, \dots, \nu+1. \tag{1.6}$$

We denote an open ball in $\mathcal{A}_m$ centered at u of radius δ by $B_\delta(u)$, i.e., $B_\delta(u) = \{v \in \mathcal{A}_m : \|v - u\|_2 < \delta\}$.

Our main result of the paper is the following theorem.

THEOREM 1.1. *Let $\Omega = (0,1)$. For each positive integer N we can find $\delta > 0$ such that*

(1) *$\{B_\delta(U_{\nu,1}), B_\delta(U_{\nu,2}) : \nu = 1, 2, \dots, N\}$ is a family of $2N$ mutually disjoint open balls in $\mathcal{A}_m$;*

(2) *there exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$, every ν, $\nu = 1, 2, \dots, N$, there exist a local minimum $u_{\epsilon,\nu,1}$ of $\mathcal{I}_\epsilon$ (or $\mathcal{J}_\epsilon$) in $B_\delta(U_{\nu,1})$ and a local minimum $u_{\epsilon,\nu,2}$ in $B_\delta(U_{\nu,2})$ satisfying $\lim_{\epsilon\to 0} \|u_{\epsilon,\nu,1} - U_{\nu,1}\|_2 = 0$ and $\lim_{\epsilon\to 0} \|u_{\epsilon,\nu,2} - U_{\nu,2}\|_2 = 0$.*

Our second result describes the global minima of $\mathcal{I}_\epsilon$ and $\mathcal{J}_\epsilon$.

THEOREM 1.2. *Let $\Omega = (0,1)$ and u_ϵ be a global minimum of $\mathcal{I}_\epsilon$ (or $\mathcal{J}_\epsilon$). There exists a countable set $\mathcal{C} \subset R$ with the following properties.*

(1) *If* $\frac{(1-m^2)^2}{24\gamma^2 c_0} \notin \mathcal{C}$, *there exists a positive integer* ν_* *so that for every* $\delta > 0$ *there is* ϵ_0 *so that if* $\epsilon < \epsilon_0$, u_ϵ *is in the open ball* $B_\delta(U_{\nu_*,1})$ *or* $B_\delta(U_{\nu_*,2})$.

(2) *If* $\frac{(1-m^2)^2}{24\gamma^2 c_0} \in \mathcal{C}$, *there exists a positive integer* ν_* *so that for every* $\delta > 0$ *there is* ϵ_0 *so that if* $\epsilon < \epsilon_0$, u_ϵ *is in one of the following four open balls:* $B_\delta(U_{\nu_*,1})$, $B_\delta(U_{\nu_*,2})$, $B_\delta(U_{\nu_*+1,1})$, *or* $B_\delta(U_{\nu_*+1,2})$.

c_0 *is given in* (2.2), *and* $\mathcal{C}$ *and* ν_* *are defined near the end of section* 3.

If we take N in Theorem 1.1 to be greater than ν_* in Theorem 1.2 and δ in Theorem 1.2 to be the same as the δ in Theorem 1.1, then every global minimum u_ϵ for small ϵ is a local minimum shown to exist in Theorem 1.1.

In the proof of Theorem 1.2 the reader will see that if the functional is $\mathcal{I}_\epsilon$, $|\nu_* - \max\{1, (\frac{(1-m^2)^2}{12\gamma^2 c_0})^{1/3}\}| < 1$. The estimate of ν_* is weaker in the case of $\mathcal{J}_\epsilon$. There ν_* is close to $(\frac{(1-m^2)^2}{12\gamma^2 c_0})^{1/3}$ only if c_0 is small.

If $W \in C^1(R)$, the local minima $u_{\epsilon,\nu,1}$, $u_{\epsilon,\nu,2}$, and the global minimum u_ϵ of $\mathcal{I}_\epsilon$, together with a v and a λ, solve the Euler equation of $\mathcal{I}_\epsilon$:

$$\begin{aligned}
&-\epsilon\Delta u + \frac{1}{\epsilon}W'(u) + v = \lambda, \ x \in \Omega,\\
&-\gamma^2\Delta v = u - m, \ x \in \Omega,\\
&\frac{\partial u}{\partial n}|_{\partial\Omega} = 0, \ \ \frac{\partial v}{\partial n}|_{\partial\Omega} = 0,\\
&\frac{1}{|\Omega|}\int_\Omega u\,dx = m, \ \frac{1}{|\Omega|}\int_\Omega v\,dx = 0.
\end{aligned}$$

For $\mathcal{J}_\epsilon$ the Euler equation is

$$\begin{aligned}
&-\epsilon\Delta u + \frac{1}{\epsilon}W'(u) - u + v = \lambda, \ x \in \Omega,\\
&-\gamma^2\Delta v + v = u, \ x \in \Omega,\\
&\frac{\partial u}{\partial n}|_{\partial\Omega} = 0, \ \frac{\partial v}{\partial n}|_{\partial\Omega} = 0,\\
&\frac{1}{|\Omega|}\int_\Omega u\,dx = m.
\end{aligned}$$

We point out that there is a large literature on the local variational problem, $\mathcal{I}_\epsilon$ without $\frac{1}{2}|(-\gamma^2\Delta)^{-1/2}(u-m)|^2$ term, and its Γ-limit. We refer to Modica [7] for the Γ-limit and Kohn and Sternberg [6] for local minimum solutions. We also refer the reader to Dal Maso [3] for the general Γ-convergence theory.

The presence of the nonlocal terms, $\frac{1}{2}|(-\gamma^2\Delta)^{-1/2}(u-m)|^2$ in $\mathcal{I}_\epsilon$ and $\frac{1}{2}|(-\gamma^2\Delta + 1)^{-1/2}u|^2$ in $\mathcal{J}_\epsilon$, gives us local minima with arbitrarily many transitional layers. This contrasts sharply with the local problem, where, according to a result of Carr, Gurtin, and Slemrod [2], every local minimum must be monotone.

We will only present the complete proof of Theorems 1.1 and 1.2 for $\mathcal{I}_\epsilon$. In section 2 we identify the limiting problem of $\mathcal{I}_\epsilon$ and show that the existence of isolated local minima of the limiting problem implies the existence of local minima of $\mathcal{I}_\epsilon$. Then in section 3 we prove that the limiting problem admits many isolated local minima, hence proving Theorem 1.1. Theorem 1.2 is also proved in that section. The study of J_ϵ is quite similar to that of $\mathcal{I}_\epsilon$. We list the modifications one needs in order to obtain the theorems for $\mathcal{J}_\epsilon$ in section 4.

2. The Γ-limit of $\mathcal{I}_\epsilon$. Associated with $\mathcal{I}_\epsilon$ is the variational problem

$$\mathcal{I}_0(u) = \begin{cases} \dfrac{c_0}{2}\|Du\|(\Omega) + \displaystyle\int_\Omega \frac{1}{2}|(-\gamma^2\Delta)^{-1/2}(u-m)|^2\,dx, \\ \qquad\qquad\qquad\qquad\qquad \text{if } u \in \mathcal{A}_m \cap BV(\Omega,\{-1,1\}), \\ \infty \qquad\qquad\qquad\qquad\quad \text{if } u \in \mathcal{A}_m \backslash BV(\Omega,\{-1,1\}) \end{cases} \tag{2.1}$$

for $u \in \mathcal{A}_m$. Here

$$c_0 = \sqrt{2}\int_{-1}^{1}(W(s))^{1/2}ds. \tag{2.2}$$

$BV(\Omega,\{-1,1\}) = \{u \in BV(\Omega) : u(x) = -1 \text{ or } 1 \text{ for almost everywhere (a.e.) } x \in \Omega\}$. $BV(\Omega)$ is the space of functions of bounded variation. We refer the reader to [4, Chap. 5, pp. 166–226] for its properties. $\|Du\|$ is the absolute value of the distributional derivative Du of u, regarded as a finite nonnegative measure on Ω. $\|Du\|(\Omega)$ is the size of Ω under this measure.

Proposition 2.1.

(1) *For every family* $\{u_\epsilon\} \subset \mathcal{A}_m$ *with* $\lim_{\epsilon\to 0} u_\epsilon = u$,

$$\liminf_{\epsilon\to 0}\mathcal{I}_\epsilon(u_\epsilon) \geq \mathcal{I}_0(u).$$

(2) *For every* $u \in \mathcal{A}_m \cap BV(\Omega,\{-1,1\})$, *there exists a family* $\{u_\epsilon\} \subset \mathcal{A}_m$ *such that* $\lim_{\epsilon\to 0} u_\epsilon = u$, *and*

$$\limsup_{\epsilon\to 0}\mathcal{I}_\epsilon(u_\epsilon) \leq \mathcal{I}_0(u).$$

Proof. We define three functionals on $\mathcal{A}_m$:

$$\mathcal{H}_\epsilon(u) = \begin{cases} \displaystyle\int_\Omega \left\{\frac{\epsilon}{2}|\nabla u|^2 + \frac{1}{\epsilon}W(u)\right\}dx & \text{if } u \in \mathcal{A}_m \cap W^{1,2}(\Omega), \\ \infty & \text{if } u \in \mathcal{A}_m \backslash W^{1,2}(\Omega), \end{cases} \tag{2.3}$$

$$\mathcal{H}_0(u) = \begin{cases} \dfrac{c_0}{2}\|Du\|(\Omega) & \text{if } u \in \mathcal{A}_m \cap BV(\Omega,\{-1,1\}), \\ \infty & \text{if } u \in \mathcal{A}_m \backslash BV(\Omega,\{-1,1\}), \end{cases} \tag{2.4}$$

and

$$\mathcal{K}(u) = \int_\Omega \frac{1}{2}[(-\gamma^2\Delta)^{-1/2}(u-m)]^2\,dx, \quad u \in \mathcal{A}_m. \tag{2.5}$$

Then $\mathcal{I}_\epsilon = \mathcal{H}_\epsilon + \mathcal{K}$ and $\mathcal{I}_0 = \mathcal{H}_0 + \mathcal{K}$. After making some minor modifications (change $L^1(\Omega)$ to $L^2(\Omega)$) in the proof of Propositions 1 and 2 in Modica [7], we find (1) for every family $\{u_\epsilon\} \subset \mathcal{A}_m$ with $\lim_{\epsilon\to 0} u_\epsilon = u$,

$$\liminf_{\epsilon\to 0}\mathcal{H}_\epsilon(u_\epsilon) \geq \mathcal{H}_0(u);$$

(2) for every $u \in \mathcal{A}_m \cap BV(\Omega,\{-1,1\})$, there exists a family $\{u_\epsilon\} \subset \mathcal{A}_m$ such that $\lim_{\epsilon\to 0} u_\epsilon = u$, and

$$\limsup_{\epsilon\to 0}\mathcal{H}_\epsilon(u_\epsilon) \leq \mathcal{H}_0(u).$$

Then we note that $\mathcal{K} : \mathcal{A}_m \to R$ is a continuous functional. Hence the two statements about $\mathcal{H}_\epsilon$ and $\mathcal{H}_0$ are carried over to $\mathcal{I}_\epsilon$ and $\mathcal{I}_0$. □

The notion of Γ-convergence is indeed defined by the two properties of Proposition 2.1. So $\mathcal{I}_\epsilon$ Γ-converges to $\mathcal{I}_0$.

Throughout the rest of this paper we assume $\Omega = (0,1)$.

PROPOSITION 2.2. *Let ϵ_n be a sequence of positive numbers converging to* 0*, and let $\{u_n\}$ be a sequence in $\mathcal{A}_m$. If $\mathcal{I}_{\epsilon_n}(u_n)$ is bounded above in n, then $\{u_n\}$ is relatively compact in $\mathcal{A}_m$ and its cluster points belong to $BV((0,1),\{-1,1\})$.*

Proof. We set

$$\phi(t) = \int_{-1}^{t} W^{1/2}(s)ds. \tag{2.6}$$

Then (1.2) implies

$$|\phi(t)| \le C + C|t|^{\frac{k}{2}+1}.$$

Set $v_n = \phi(u_n)$. As shown in Modica and Mortola [8], v_n is bounded in $W^{1,1}(0,1)$. For by (1.2) and $k \ge 2$, we find

$$|v_n| \le C + C|u_n|^{\frac{k}{2}+1} \le C + CW(u_n).$$

Therefore $\{v_n\}$ is bounded in $L^1(0,1)$. On the other hand,

$$\int_0^1 |v_n'|\, dx = \int_0^1 W^{1/2}(u_n)|u_n'|\, dx$$

$$\le \frac{\sqrt{2}}{2}\left(\int_0^1 \left[\frac{\epsilon_n}{2}|u_n'|^2 + \frac{1}{\epsilon_n}W(u_n)\right] dx\right) \le \frac{\sqrt{2}}{2}\mathcal{I}_{\epsilon_n}(u_n).$$

Therefore $\{v_n\}$ is bounded in $W^{1,1}(0,1)$. The Sobolev imbedding theorem asserts that $\{v_n\}$ is relatively compact in $L^p(0,1)$ for all $1 \le p < \infty$.

Now consider $u_n = \phi^{-1}(v_n)$. (1.2) and (2.6) imply that ϕ^{-1} is continuous and increasing, and there exists $\bar{t} > 0$ such that ϕ^{-1} is Lipschitz continuous for $|t| \ge \bar{t}$. We can find C such that for all t

$$|\phi^{-1}(t)| \le C + C|t|, \ |\phi^{-1}(t)|^p \le C + C|t|^p. \tag{2.7}$$

To prove that $\{u_n\}$ is relatively compact we show that every subsequence of $\{u_n\}$ has an L^2-convergent further subsequence. Let us recall Vitali's convergence theorem [5, p. 203].

Vitali's convergence theorem. *Let $\{f_n\}$ be a sequence in $L^p(\Omega,\mu)$, $1 \le p < \infty$, and f be an μ-measurable function such that $f_n \to f$ μ-a.e.. Then $f \in L^p(\Omega,\mu)$ and $\|f_n - f\|_p \to 0$ if and only if*

1. *for each $\varepsilon > 0$, there exists a μ-measurable set $A_\varepsilon \subset \Omega$ such that $\mu(A_\varepsilon) < \infty$ and $\int_{\Omega\setminus A_\varepsilon} |f_n|^p\, d\mu < \varepsilon$ for all n; and*
2. *for each $\varepsilon > 0$, there is $\delta > 0$ such that for every μ measurable set E, $\mu(E) < \delta$ implies $\int_E |f_n|^p\, d\mu < \varepsilon$ for all n.*

Part 1 of Vitali's convergence theorem is not needed here because $(0,1)$ itself has finite Lebesgue measure. Let $\{u_{n_l}\}$ be a subsequence of $\{u_n\}$. Then there are a subsequence of $\{v_{n_l} = \phi(u_{n_l})\}$, denoted by $\{v_{n_{l_m}}\}$, and $v \in L^p(0,1)$ such that $v_{n_{l_m}} \to v$ in $L^p(0,1)$ and $v_{n_{l_m}} \to v$ a.e. Then $u_{n_{l_m}} \to \phi^{-1}(v)$ a.e. Applying Vitali's convergence theorem to $v_{n_{l_m}}$, we find that for every $\varepsilon > 0$ there is $\delta > 0$ such that for every measurable set E, $|E| < \delta$ implies $\int_E |v_{n_{l_m}}|^p \, dx < \varepsilon$ for all n. Then (2.7) implies

$$\int_E |u_{n_{l_m}}|^p \, dx \le \int_E (C + C|v_{n_{l_m}}|^p) \, dx < C\delta + C\varepsilon.$$

Now Vitali's convergence theorem applied to $\{u_{n_{l_m}}\}$ asserts that $u_{n_{l_m}} \to \phi^{-1}(v)$ in $L^p(0,1)$, particularly in $L^2(0,1)$.

Let u be a cluster point of $\{u_n\}$, i.e., there exists a subsequence $\{u_{n_l}\}$ such that $u_{n_l} \to u$ in $L^2(0,1)$ as $l \to \infty$. Fatou's lemma and the boundedness of $\mathcal{I}_{\epsilon_n}(u_n)$ imply that

$$0 \le \int_0^1 W(u) \le \lim_{l\to\infty} \int_0^1 W(u_{n_l}) \, dx \le \lim_{l\to\infty} \epsilon_{n_l} \mathcal{I}_{\epsilon_{n_l}}(u_{n_l}) = 0.$$

Then for a.e. $x \in (0,1)$, $u(x) = -1$ or 1. If we consider $v_{n_l} = \phi(u_{n_l})$, then the boundedness of $\{v_{n_l}\}$ in $W^{1,1}(0,1)$, proved earlier, implies that $\phi(u)$, the L^1-limit of $\{v_{n_l}\}$, is a BV function [4, Thm. 1, p. 172]. $\phi(u)$ only takes two values, $\phi(-1)$ and $\phi(1)$. Then $\phi(u) = \phi(-1) + \frac{\phi(1)-\phi(-1)}{2}(u+1)$. Hence u is also a BV function. $\square$

A useful property following Propositions 2.1 and 2.2 is that isolated local minima of the Γ-limit persist under small perturbation. It was used in Kohn and Sternberg [6], (see also Dal Maso [3]). We include this property and its proof below for completeness.

PROPOSITION 2.3. *Let $\delta > 0$ and $u_0 \in \mathcal{A}_m$ be such that $\mathcal{I}_0(u_0) < \mathcal{I}_0(u)$ for all $u \in B_\delta(u_0)$ with $u \ne u_0$. Then there exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$ there exists $u_\epsilon \in B_{\delta/2}(u_0)$ with $\mathcal{I}_\epsilon(u_\epsilon) \le \mathcal{I}_\epsilon(u)$ for all $u \in B_{\delta/2}(u_0)$. In addition* $\lim_{\epsilon\to 0} \|u_\epsilon - u_0\|_2 = 0$.

Proof. Let $u_{\epsilon,n}$ be a sequence in $B_{\delta/2}(u_0)$ so that

$$\lim_{n\to\infty} \mathcal{I}_\epsilon(u_{\epsilon,n}) = \inf_{u\in B_{\delta/2}(u_0)} \mathcal{I}_\epsilon(u).$$

The standard argument shows that after passing to a subsequence, again denoted by $u_{\epsilon,n}$, there exists $u_\epsilon \in \overline{B}_{\delta/2}(u_0)$ such that $u_{\epsilon,n} \to u_\epsilon$ in $L^2(0,1)$, $u_{\epsilon,n} \to u_\epsilon$ weakly in $W^{1,2}(0,1)$, and

$$\mathcal{I}_\epsilon(u_\epsilon) = \lim_{n\to\infty} \mathcal{I}_\epsilon(u_{\epsilon,n}) = \inf_{u\in B_{\delta/2}(u_0)} \mathcal{I}_\epsilon(u).$$

Next we claim $u_\epsilon \in B_{\delta/2}(u_0)$ if ϵ is small enough. Otherwise there exists a sequence $\epsilon_l \to 0$, such that $\|u_{\epsilon_l} - u_0\|_2 = \delta/2$ and

$$\mathcal{I}_{\epsilon_l}(u_{\epsilon_l}) = \inf_{u\in B_{\delta/2}(u_0)} \mathcal{I}_{\epsilon_l}(u).$$

Part 2 of Proposition 2.1 asserts that there exists a sequence v_{ϵ_l} in $B_{\delta/2}(u_0)$, if l is large enough, such that

$$\limsup_{l\to\infty} \mathcal{I}_{\epsilon_l}(v_{\epsilon_l}) \le \mathcal{I}_0(u_0).$$

Therefore,

$$\limsup_{l\to\infty} \mathcal{I}_{\epsilon_l}(u_{\epsilon_l}) \le \limsup_{l\to\infty} \mathcal{I}_{\epsilon_l}(v_{\epsilon_l}) \le \mathcal{I}_0(u_0).$$

Proposition 2.2 then asserts that, after passing to a subsequence, again denoted by u_{ϵ_l}, there exists $\overline{u}_0$ such that $u_{\epsilon_l} \to \overline{u}_0$ in $L^2(0,1)$, and $\|\overline{u}_0 - u_0\|_2 = \delta/2$. Part 1 of Proposition 2.1 now implies

$$\mathcal{I}_0(\overline{u}_0) \le \liminf_{l\to\infty} \mathcal{I}_{\epsilon_l}(u_{\epsilon_l}) \le \mathcal{I}_0(u_0).$$

This contradicts the condition that $\mathcal{I}_0(u_0) < \mathcal{I}_0(u)$ for all $u \in B_\delta(u_0)$ with $u \ne u_0$. Therefore u_ϵ is in the open ball $B_{\delta/2}(u_0)$, i.e., u_ϵ is a local minimum.

To show $u_\epsilon \to u_0$ in $L^2(0,1)$ as $\epsilon \to 0$, we assume that there exists a sequence $\epsilon_l \to 0$ such that $\|u_{\epsilon_l} - u_0\|_2 = \delta_0 < \delta/2$. Then arguing like above, we have $\widetilde{u}_0$ such that, after passing to a subsequence, again denoted by u_{ϵ_l}, $u_{\epsilon_l} \to \widetilde{u}_0$ and $\|\widetilde{u}_0 - u_0\|_2 = \delta_0$. Then by part 1 of Proposition 2.1,

$$\mathcal{I}_0(\widetilde{u}_0) \le \liminf_{l\to\infty} \mathcal{I}_{\epsilon_l}(u_{\epsilon_l}) \le \mathcal{I}_0(u_0),$$

which is again a contradiction. □

3. The local minima of $\mathcal{I}_0$.

A function u in $BV((0,1),\{-1,1\})$, up to a set of Lebesgue measure 0, is a step function. $u(x)$ switches between -1 and 1 at finitely many points $x_1, x_2, \dots, x_\nu$, with $0 < x_1 < x_2 < \cdots < x_\nu < 0$. A formal description is as follows.

For $u \in BV((0,1),\{-1,1\})$ we define set $E_u = \{x \in (0,1) : u(x) = -1\}$. The perimeter of E_u in $(0,1)$ is $\|D\chi_{E_u}\|(0,1)$, where χ_{E_u} is the characteristic function of E_u. The measure $\|D\chi_{E_u}\|$ is often written as $\|\partial E_u\|$. Clearly $\|\partial E_u\| = \frac{\|Du\|}{2}$.

The reduced boundary of E_u, a subset of $(0,1)$, is denoted by $\partial^* E_u$ (see [4, section 5.7, pp. 194–207] for the definition and properties of reduced boundaries). The structure theorem for sets of finite perimeter [4, Thm. 1 (iii), p. 189] asserts that $\|\partial E_u\| = H^0 | \partial^* E_u$, the 0-dimensional Hausdorff measure restricted on $\partial^* E_u$. The 0-dimensional Hausdorff measure is the counting measure. Therefore $\partial^* E_u$ is a set of finitely many points in $(0,1)$ and $\|\partial E_u\|(0,1) = \frac{\|Du\|(0,1)}{2}$ is the number of the points in $\partial^* E_u$.

$\partial^* E_u$ is simply $\{x_1, x_2, \dots, x_\nu\}$, the set of points where $u(x)$ switches, and

$$\frac{\|Du\|(0,1)}{2} = \nu.$$

If $u \in \mathcal{A}_m \cap BV((0,1),\{-1,1\})$, $\frac{\|Du\|(0,1)}{2}$ has to be nonzero. Otherwise u would be a constant. Then $u = -1$ for a.e. $x \in (0,1)$ or $u = 1$ for a.e. $x \in (0,1)$. In either case $\int_0^1 u \ne m$. So we have the following mutually disjoint decomposition:

$$\begin{aligned} &\mathcal{A}_m \cap BV((0,1),\{-1,1\}) = \cup_1^\infty A_\nu, \text{ where} \\ &A_\nu = \left\{ u \in \mathcal{A}_m \cap BV((0,1),\{-1,1\}) : \frac{\|Du\|(0,1)}{2} = \nu \right\}. \end{aligned} \tag{3.1}$$

Proposition 3.1. *For every $u \in A_\nu$, $u \ne U_{\nu,1}$, and $u \ne U_{\nu,2}$, we have $\mathcal{I}_0(U_{\nu,1}) = \mathcal{I}_0(U_{\nu,2}) < \mathcal{I}_0(u)$.*

Proof. For each $u \in A_\nu$, let us denote $\partial^* E_u$ by $\{x_1, x_2, \ldots, x_\nu\}$, where $0 < x_1 < x_2 \cdots < x_\nu < 1$. Since $\|Du\|(x_i, x_{i+1}) = 0$ for each i and (x_i, x_{i+1}) is connected, $u = -1$ for a.e. $x \in (x_i, x_{i+1})$ or $u = 1$ for a.e. $x \in (x_i, x_{i+1})$. And it follows from the definition of reduced boundaries [4, p. 194] that $u(x)$ must jump from -1 to 1 or 1 to -1 when x moves from (x_{i-1}, x_i) to (x_i, x_{i+1}). We can further decompose A_ν into two disjoint sets:

$$\begin{aligned} A_{\nu,1} &= \{u \in A_\nu : u = -1 \text{ for a.e. } x \in (0, x_1)\}, \\ A_{\nu,2} &= \{u \in A_\nu : u = 1 \text{ for a.e. } x \in (0, x_1)\}. \end{aligned} \tag{3.2}$$

For $u \in A_{\nu,1}$ the constraint $\int_0^1 u = m$ becomes $-2x_1 + 2x_2 - \cdots + 2(-1)^\nu x_\nu - (-1)^\nu = m$, and for $u \in A_{\nu,2}$ the constraint $\int_0^1 u = m$ becomes $2x_1 - 2x_2 - \cdots - 2(-1)^\nu x_\nu + (-1)^\nu = m$.

Now $A_{\nu,1}$ can be identified with the set

$$\begin{aligned} S_{\nu,1} = \Big\{&(x_1, \ldots, x_\nu) \in R^\nu : 0 < x_1 < \cdots x_\nu < 1, \\ &-x_1 + x_2 - \cdots + (-1)^\nu x_\nu = \frac{m + (-1)^\nu}{2}\Big\}, \end{aligned} \tag{3.3}$$

and $A_{\nu,2}$ can be identified with the set

$$\begin{aligned} S_{\nu,2} = \Big\{&(x_1, \ldots, x_\nu) \in R^\nu : 0 < x_1 < \cdots x_\nu < 1, \\ &x_1 - x_2 - \cdots - (-1)^\nu x_\nu = \frac{m - (-1)^\nu}{2}\Big\}. \end{aligned} \tag{3.4}$$

$S_{\nu,1}$ and $S_{\nu,2}$ are two bounded open subsets of two $\nu - 1$ dimensional hyper-planes of R^ν.

We take $u \in A_{\nu,1} \cong S_{\nu,1}$ and $\{x_1, x_2, \ldots, x_\nu\}$, $x_1 < \cdots < x_\nu$, to be $\partial^* E_u$. We compute $\mathcal{K}(u)$. Let v be the solution of

$$-\gamma^2 v'' = u - m, \ v'(0) = v'(1) = 0, \ \int_0^1 v = 0.$$

Denote the Green's function of this equation by $G(x, y)$. Then

$$\begin{aligned} \mathcal{K}(u) &= \frac{1}{2} \int_0^1 [(-\gamma^2 \Delta)^{-1/2} (u - m)]^2 \, dx \\ &= \frac{1}{2} \int_0^1 (u - m) v \, dx \\ &= \frac{1}{2} \left[\int_0^{x_1} (-1 - m) v + \int_{x_1}^{x_2} (1 - m) v + \cdots + \int_{x_\nu}^1 ((-1)^{\nu+1} - m) v \right]. \end{aligned}$$

Treating $\mathcal{K}$ as a function of $(x_1, x_2, \ldots, x_\nu)$ in $S_{\nu,1}$, we calculate

$$\frac{\partial \mathcal{K}}{\partial x_1} = \frac{1}{2} \left[-2v(x_1) + \int_0^1 (u - m) \frac{\partial v}{\partial x_1} \, dx \right].$$

Since

$$\begin{aligned} \frac{\partial v}{\partial x_1}(x) = \frac{\partial}{\partial x_1} \bigg[&\int_0^{x_1} (-1 - m) G(x, y) dy + \int_{x_1}^{x_2} (1 - m) G(x, y) dy + \cdots \\ &+ \int_{x_\nu}^1 ((-1)^{\nu+1} - m) G(x, y) dy \bigg] \\ = -2G(x, x_1),& \end{aligned}$$

we deduce

$$\begin{aligned}\frac{\partial \mathcal{K}}{\partial x_1} &= \frac{1}{2}\left[-2v(x_1) + \int_0^1 (u-m)(-2)G(x,x_1)\,dx\right]\\ &= -2v(x_1).\end{aligned}$$

The same argument applied to other x_i yields

$$\nabla \mathcal{K} = 2(-v(x_1), v(x_2), \dots, (-1)^\nu v(x_\nu)). \tag{3.5}$$

Since $\int_0^1 u = m$, or $-x_1 + x_2 - \dots + (-1)^\nu x_\nu = \frac{m+(-1)^\nu}{2}$, the Lagrange multiplier method asserts that if $(x_1, x_2, \dots, x_\nu)$ is a critical point of $\mathcal{K}$ in $S_{\nu,1}$, there exists λ such that

$$\nabla \mathcal{K} = \lambda(-1, 1, -1, \dots, (-1)^\nu),$$

which, together with (3.5), implies

$$v(x_1) = v(x_2) = \dots = v(x_\nu). \tag{3.6}$$

On (x_1, x_2), v solves the linear equation $-\gamma^2 v'' = 1 - m$. Then $v(x_1) = v(x_2)$ implies that v is symmetric about $(x_1 + x_2)/2$, and hence $v'(x_1) = -v'(x_2)$. On intervals $(0, x_1)$ and (x_2, x_3), v satisfies the linear equation $-\gamma^2 v'' = -1 - m$. Since v also satisfies the conditions $v(x_1) = v(x_2)$, $v'(x_1) = -v'(x_2)$, $v(x_2) = v(x_3)$, and $v'(0) = 0$, we conclude by solving the equation on $(0, x_1)$ and (x_2, x_3) that v on $(0, x_1)$ is a reflection of v on $(x_2, (x_2 + x_3)/2)$. Hence the length of $(0, x_1)$ is half that of (x_2, x_3). In the next step we compare intervals (x_2, x_3) and (x_4, x_5) and similarly find that they have the same length. By repeating this argument we conclude that the intervals where $u = -1$ all have the same length with the exception of $(0, x_1)$ and $(x_\nu, 1)$ if $u = -1$ there, whose length is half. The same can be said for the intervals where $u = 1$. Taking $-x_1 + x_2 - \dots + (-1)^\nu x_\nu = \frac{m+(-1)^\nu}{2}$ into consideration, we find

$$x_1 = \frac{1-m}{2\nu}, x_2 = x_1 + \frac{1+m}{\nu}, \dots, x_\nu = x_{\nu-1} + \frac{1+(-1)^\nu m}{\nu}.$$

We have proved that $\mathcal{K}$ has a unique critical point $(x_1, x_2, \dots, x_\nu)$ in $S_{\nu,1}$. We denote the function in $A_{\nu,1}$ whose reduced boundary is $\{x_1, x_2, \dots, x_\nu\}$ by $U_{\nu,1}$ (also defined in (1.5)). We proceed to prove that $U_{\nu,1}$ minimizes $\mathcal{K}$ in $A_{\nu,1}$. We first compute $\mathcal{K}(U_{\nu,1})$. Let v be the solution of

$$-\gamma^2 v'' = U_{\nu,1} - m,\ v'(0) = v'(1) = 0,\ \int_0^1 v = 0.$$

Then

$$\mathcal{K}(U_{\nu,1}) = \frac{1}{2}\int_0^1 (U_{\nu,1} - m)v = \frac{\gamma^2}{2}\int_0^1 |v'|^2.$$

On $(0, x_1)$ $v'(x) = \frac{-1-m}{-\gamma^2}x + v'(0) = \frac{-1-m}{-\gamma^2}x$. Then

$$\frac{\gamma^2}{2}\int_0^{x_1} |v'|^2 = \frac{\gamma^2}{2}\frac{-\gamma^2}{3(-1-m)}(v'(x_1))^3.$$

On (x_1, x_2) $v'(x) = \frac{1-m}{-\gamma^2}(x - x_1) + v'(x_1)$. Then

$$\frac{\gamma^2}{2}\int_{x_1}^{x_2} |v'|^2 = \frac{\gamma^2}{2}\frac{-\gamma^2}{3(1-m)}[(v'(x_2))^3 - (v'(x_1))^3].$$

After finding $\frac{\gamma^2}{2}\int_{x_i}^{x_{i+1}} |v'|^2$ on each (x_i, x_{i+1}) and summing over i, we deduce

$$\mathcal{K}(U_{\nu,1}) = \frac{\gamma^4}{3(1-m^2)}[(v'(x_1))^3 - (v'(x_2))^3 + (v'(x_3))^3 + \cdots + (-1)^{\nu+1}(v'(x_\nu))^3].$$

Since we also know $v'(x_1) = -v'(x_2) = v'(x_3) = \cdots = (-1)^{\nu+1}v'(x_\nu)$ and $v'(x_1) = \frac{-1-m}{-\gamma^2}\frac{1-m}{2\nu} = \frac{1-m^2}{2\gamma^2\nu}$, we find

$$\mathcal{K}(U_{\nu,1}) = \frac{(1-m^2)^2}{24\gamma^2\nu^2}. \tag{3.7}$$

A similar computation in $A_{\nu,2}$ finds $U_{\nu,2}$ (defined by (1.6)) and again

$$\mathcal{K}(U_{\nu,2}) = \frac{(1-m^2)^2}{24\gamma^2\nu^2}. \tag{3.8}$$

We now show that $\mathcal{K}(u) > \mathcal{K}(U_{\nu,1})$ for every $u \in A_{\nu,1} \cong S_{\nu,1}$, $u \neq U_{\nu,1}$. If this is not the case, since there is only one critical point, $U_{\nu,1}$, in $S_{\nu,1}$, there must be a sequence $\{(x_{n,1}, x_{n,2}, \ldots, x_{n,\nu})\}$ converging to a point $(y_1, y_2, \ldots, y_\nu)$ on the boundary of $S_{\nu,1}$ ($S_{\nu,1}$ is considered as a subset of R^ν) such that

$$\lim_{n\to\infty} \mathcal{K}(x_{n,1}, x_{n,2}, \ldots, x_{n,\nu}) \leq \mathcal{K}(U_{\nu,1}).$$

For the point $(y_1, y_2, \ldots, y_\nu)$ to be on the boundary of $S_{\nu,1}$, at least two of 0, $y_1, \ldots, y_\nu$, 1 must be identical. Then $(y_1, y_2, \ldots, y_\nu)$ is identified as a point in $S_{\nu',1}$ or $S_{\nu',2}$, corresponding to $A_{\nu',1}$ or $A_{\nu',2}$ for some $\nu' < \nu$. Let us denote this point by $(z_1, z_2, \ldots, z_{\nu'})$ and assume, without the loss of generality, $(z_1, z_2, \ldots, z_{\nu'}) \in S_{\nu',1}$. We ask whether $U_{\nu',1}$ is the strict global minimum of $\mathcal{K}$ in $A_{\nu',1}$. If so,

$$\mathcal{K}(U_{\nu',1}) \leq \mathcal{K}(z_1, z_2, \ldots, z_{\nu'}) = \lim_{n\to\infty} \mathcal{K}(x_{n,1}, x_{n,2}, \ldots, x_{n,\nu}) \leq \mathcal{K}(U_{\nu,1}),$$

which, since $\nu' < \nu$, is inconsistent with (3.7), where $\mathcal{K}(U_{\nu,1}) = \mathcal{K}(U_{\nu,2})$ decreases in ν. If $U_{\nu',1}$ is not the strict global minimum of $\mathcal{K}$ in $A_{\nu',1}$, we use the same argument on $U_{\nu',1}$ and end up in a $A_{\nu'',1}$ or $A_{\nu'',2}$ with $\nu'' < \nu'$. This process stops at $\nu = 1$, and there, since $A_{1,1}$ has only one element $U_{1,1}$ and $A_{1,2}$ has only one element $U_{1,2}$, we find $\mathcal{K}(U_{1,1}) = \mathcal{K}(U_{1,2}) \leq \mathcal{K}(U_{\nu,1})$, which is inconsistent with (3.7) or (3.8). So we have proved that $U_{\nu,1}$ is the strict global minimum of $\mathcal{K}$ in $A_{\nu,1}$. And since for $u \in A_{\nu,1}$, $\mathcal{I}_0(u) = c_0\nu + \mathcal{K}(u)$, Proposition 3.1 is proved. □

PROPOSITION 3.2. *Let N be a positive integer. One can find $\delta > 0$ such that*

(1) *$\{B_\delta(U_{\nu,1}), B_\delta(U_{\nu,2}) : \nu = 1, 2, \ldots, N\}$ is a family of $2N$ mutually disjoint open balls in $\mathcal{A}_m$;*

(2) *for all $u \in B_\delta(U_{\nu,1})$, $\nu = 1, 2, \ldots, N$, with $u \neq U_{\nu,1}$, $\mathcal{I}_0(U_{\nu,1}) < \mathcal{I}_0(u)$, and for all $u \in B_\delta(U_{\nu,2})$ with $u \neq U_{\nu,2}$, $\mathcal{I}_0(U_{\nu,2}) < \mathcal{I}_0(u)$.*

Proof. Let N be a positive integer and $\nu \in \{1, 2, \ldots, N\}$. We consider $U_{\nu,1}$. The study of $U_{\nu,2}$ is the same.

Take δ to be a positive number to be specified later. Let $u \in B_\delta(U_{\nu,1})$, $u \neq U_{\nu,1}$. If $u \in A_\nu$, then Proposition 3.1 implies Proposition 3.2. So we assume $u \in \mathcal{A}_m \backslash A_\nu$. Then if $u \in (\mathcal{A}_m \backslash A_\nu) \backslash BV((0,1),\{-1,1\})$, $\mathcal{I}_0(U_{\nu,1}) < \mathcal{I}_0(u) = \infty$. So we need only to consider $u \in (\mathcal{A}_m \backslash A_\nu) \cap BV((0,1),\{-1,1\})$.

Because of (3.1), the positive integer $\frac{\|Du\|(0,1)}{2}$ is either $\leq \nu - 1$ or $\geq \nu + 1$. We study these two cases separately.

First we prove that the case $\frac{\|Du\|(0,1)}{2} \leq \nu - 1$ does not happen if δ is small enough. We claim that there is $\delta > 0$ such that for all $u \in B_\delta(U_{\nu,1}) \cap BV((0,1),\{-1,1\})$, $\frac{\|Du_n\|(0,1)}{2} \geq \nu$. Otherwise there exist $\delta_n \to 0$ and $u_n \in B_{\delta_n}(U_{\nu,1}) \cap BV((0,1),\{-1,1\})$ such that $\frac{\|Du\|(0,1)}{2} \leq \nu - 1$. Then $u_n \to U_{\nu,1}$ in $L^2(0,1)$ implies (see [4, Thm. 1, p. 172])

$$2\nu = \|DU_{\nu,1}(0,1)\| \leq \liminf_{n\to\infty} \|Du_n\|(0,1) \leq 2(\nu - 1),$$

a contradiction.

Second we consider the case $\frac{\|Du\|(0,1)}{2} \geq \nu + 1$. Here

$$\begin{aligned}\mathcal{I}_0(u) &\geq c_0(\nu+1) + \int_0^1 \frac{1}{2}[(-\gamma^2\Delta)^{-1/2}(u-m)]^2\,dx \\ &= \mathcal{I}_0(U_{\nu,1}) + c_0 + \mathcal{K}(u) - \mathcal{K}(U_{\nu,1}).\end{aligned}$$

Denote the norm of bounded linear operator $(-\gamma\Delta)^{-1/2}$ from $\{u \in L^2(0,1) : \int_0^1 u = 0\}$ to $L^2(0,1)$ by $\|(-\gamma\Delta)^{-1/2}\|$. Estimate

$$\begin{aligned}&|\mathcal{K}(u) - \mathcal{K}(U_{\nu,1})| \\ &= \frac{1}{2}|\, \|(-\gamma\Delta)^{-1/2}(u-m)\|_2^2 - \|(-\gamma\Delta)^{-1/2}(U_{\nu,1}-m)\|_2^2\,| \\ &= \frac{1}{2}|\, \|(-\gamma\Delta)^{-1/2}(u-m)\|_2 - \|(-\gamma\Delta)^{-1/2}(U_{\nu,1}-m)\|_2\,| \\ &\quad \cdot \{\|(-\gamma\Delta)^{-1/2}(u-m)\|_2 + \|(-\gamma\Delta)^{-1/2}(U_{\nu,1}-m)\|_2\} \\ &\leq \frac{1}{2}\|(-\gamma\Delta)^{-1/2}(u-U_{\nu,1})\|_2 \\ &\quad \cdot \{\|(-\gamma\Delta)^{-1/2}(u-U_{\nu,1})\|_2 + 2\|(-\gamma\Delta)^{-1/2}(U_{\nu,m}-m)\|_2\} \\ &\leq \frac{\|(-\gamma\Delta)^{-1/2}\|^2}{2}\|u-U_{\nu,1}\|_2\{\|u-U_{\nu,1}\|_2 + 2\|U_{\nu,m}\|_2 + 2\|m\|_2\} \\ &\leq \frac{\|(-\gamma\Delta)^{-1/2}\|^2}{2}\delta\{\delta + 2 + 2m\}.\end{aligned}$$

We obtain, choosing δ sufficiently small,

$$\mathcal{I}_0(u) \geq \mathcal{I}_0(U_{\nu,1}) + c_0 - \frac{\|(-\gamma\Delta)^{-1/2}\|^2}{2}\delta\{\delta + 2 + 2m\} > \mathcal{I}_0(U_{\nu,1}).$$

Since there are only finitely many ν, we can choose δ so that it is independent of ν. We can also make $\{B_\delta(U_{\nu,1}), B_\delta(U_{\nu,2}) : \nu = 1, 2, \ldots, N\}$ mutually disjoint by having δ small enough. □

PROPOSITION 3.3. *$\{U_{\nu,1}, U_{\nu,2} : \nu = 1, 2, 3 \ldots\}$ is the set of all local minima of $\mathcal{I}_0$ (or $\mathcal{J}_0$).*

Proof. According to Proposition 3.2, each element in $\{U_{\nu,1}, U_{\nu,2} : \nu = 1, 2, \ldots\}$ is a local minimum of $\mathcal{I}_0$. On the other hand, if u is a local minimum of $\mathcal{I}_0$, it must be in $A_{\nu,i}$ for some $i = 1$ or 2 and $\nu = 1, 2, \ldots$. And in each $A_{\nu,i} \cong S_{\nu,i}$ there is only one critical point $U_{\nu,i}$ of $\mathcal{I}_0$ considered as a function on $S_{\nu,i}$. Then $u = U_{\nu,i}$. □

Theorem 1.1 follows immediately from Propositions 2.3 and 3.2. To prove Theorem 1.2, we note from (3.7) and (3.8) that

$$\mathcal{I}_0(U_{\nu,1}) = \mathcal{I}_0(U_{\nu,2}) = c_0\nu + \frac{(1-m^2)^2}{24\gamma^2\nu^2}. \tag{3.9}$$

Set

$$g(t) = c_0 t + \frac{(1-m^2)^2}{24\gamma^2 t^2}$$

for $t \geq 1$. Denote the global minimum of this convex function by t_0:

$$t_0 = \max\left\{1, \left(\frac{(1-m^2)^2}{12\gamma^2 c_0}\right)^{1/3}\right\}.$$

Let $[t_0]$ be the greatest integer less than or equal to t_0. Compare $g([t_0])$ and $g([t_0+1])$. Depending on $\frac{(1-m^2)^2}{24\gamma^2 c_0}$, we have $g([t_0]) = g([t_0+1])$ or $g([t_0]) \neq g([t_0+1])$. Let $\mathcal{C} \subset R$ be the set so that when $\frac{(1-m^2)^2}{24\gamma^2 c_0} \in \mathcal{C}$, $g([t_0]) = g([t_0+1])$. In this case we set $\nu_* = [t_0]$. And if $\frac{(1-m^2)^2}{24\gamma^2 c_0} \notin \mathcal{C}$, i.e., $g([t_0]) \neq g([t_0+1])$, we set

$$\nu_* = \begin{cases} [t_0] & \text{if } g([t_0]) < g([t_0]+1), \\ [t_0]+1 & \text{if } g([t_0]) > g([t_0]+1). \end{cases}$$

So if $\frac{(1-m^2)^2}{24\gamma^2 c_0} \notin \mathcal{C}$, for every $\nu \neq \nu_*$,

$$\mathcal{I}_0(U_{\nu_*,1}) = \mathcal{I}_0(U_{\nu_*,2}) < \mathcal{I}_0(U_{\nu,1}) = \mathcal{I}_0(U_{\nu,2}),$$

and if $\frac{(1-m^2)^2}{24\gamma^2 c_0} \in \mathcal{C}$, for every ν, $\nu \neq \nu_*$, and $\nu \neq \nu_* + 1$,

$$\mathcal{I}_0(U_{\nu_*,1}) = \mathcal{I}_0(U_{\nu_*,2}) = \mathcal{I}_0(U_{\nu_*+1,1}) = \mathcal{I}_0(U_{\nu_*+1,2}) < \mathcal{I}_0(U_{\nu,1}) = \mathcal{I}_0(U_{\nu,2}).$$

The case $\frac{(1-m^2)^2}{24\gamma^2 c_0} \notin \mathcal{C}$ is generic. $\mathcal{C}$ is a countable set.

Let u_ϵ be a global minimum of $\mathcal{I}_\epsilon$. The existence of u_ϵ follows from the standard argument as in the proof of Proposition 2.3. Propositions 2.1 and 2.2 imply that every sequence $\{u_{\epsilon_n}\}$ of u_ϵ has a convergent subsequence, and if u_0 is the limit of a subsequence, u_0 must be a global minimum of $\mathcal{I}_0$ in $\mathcal{A}_m \cap BV((0,1), \{-1,1\})$. The decomposition (3.1) and Proposition 3.1 imply that $u_0 = U_{\nu,1}$ or $U_{\nu,2}$ for some ν. And by the definition of ν_*, $\nu = \nu_*$ if $\frac{(1-m^2)^2}{24\gamma^2 c_0} \notin \mathcal{C}$ and $\nu = \nu_*$ or ν_*+1 if $\frac{(1-m^2)^2}{24\gamma^2 c_0} \in \mathcal{C}$. This proves Theorem 1.2.

4. The study of $\mathcal{J}_\epsilon$.

All results in sections 2 and 3 are valid for $\mathcal{J}_\epsilon$. In this section we indicate the modifications needed to prove these results for $\mathcal{J}_\epsilon$.

The condition (1.2) implies that for all small ϵ, $\mathcal{J}_\epsilon$ is bounded from below. The lower bound can be made independent of small ϵ.

The Γ-limit of $\mathcal{J}_\epsilon$ is

$$\mathcal{J}_0(u) = \begin{cases} \dfrac{c_0}{2}\|Du\|(\Omega) - \dfrac{|\Omega|}{2} + \displaystyle\int_\Omega \frac{1}{2}|(-\gamma^2\Delta+1)^{-1/2}u|^2\,dx, & \\ \qquad\qquad\qquad\qquad \text{if } u \in \mathcal{A}_m \cap BV(\Omega, \{-1,1\}), & \\ \infty \qquad\qquad\qquad\quad \text{if } u \in \mathcal{A}_m \backslash BV(\Omega, \{-1,1\}) & \end{cases} \tag{4.1}$$

for $u \in \mathcal{A}_m$. Define

$$\mathcal{L}(u) = \int_\Omega \frac{1}{2}\{-u^2 + [(-\gamma^2\Delta + 1)^{-1/2}u]^2\}\, dx \tag{4.2}$$

for $u \in \mathcal{A}_m$. Thus $\mathcal{J}_\epsilon = \mathcal{H}_\epsilon + \mathcal{L}$ and $\mathcal{J}_0 = \mathcal{H}_0 + \mathcal{L}$.

As in the proof of Proposition 3.1, we study $\mathcal{L}(u)$ for $u \in A_{\nu,1} \cong S_{\nu,1}$. Let $\{x_1, x_2, \dots, x_\nu\}$ be $\partial^* E_u$. Let v be the solution of

$$-\gamma^2 v'' + v = u,\ v'(0) = v'(1) = 0,$$

and let $G(x, y)$ be the Green's function of this equation. Then

$$\begin{aligned}\mathcal{L}(u) &= \frac{1}{2}\int_0^1 (-u^2 + uv)\\ &= \frac{1}{2}\left[-1 + \int_0^{x_1}(-1)v + \int_{x_1}^{x_2} v + \cdots + \int_{x_\nu}^1 (-1)^{\nu+1} v\right].\end{aligned}$$

We then find, treating $\mathcal{L}$ as a function of $x_1, x_2, \dots, x_\nu$,

$$\nabla\mathcal{L} = 2(-v(x_1), v(x_2), \dots, (-1)^\nu v(x_\nu)).$$

Again we see that at a critical point $(x_1, \dots, x_\nu)$, $v(x_1) = v(x_2) = \cdots = v(x_\nu)$. The same symmetry argument as in the proof of Proposition 3.1 shows that

$$x_1 = \frac{1-m}{2\nu}, x_2 = x_1 + \frac{1+m}{\nu}, \dots, x_\nu = x_{\nu-1} + \frac{1 + (-1)^\nu m}{\nu}.$$

We again obtain $U_{\nu,1}$.

The calculation of $\mathcal{L}(U_{\nu,1})$ is a bit more complex. Let v be the solution of

$$-\gamma^2 v'' + v = U_{\nu,1},\ v'(0) = v'(1) = 0.$$

Then

$$\mathcal{L}(U_{\nu,1}) = -\frac{1}{2} + \frac{1}{2}\int_0^1 U_{\nu,1} v\, dx.$$

On an subinterval of $(0, 1)$, say (a,b), where $U_{\nu,1} = \alpha \in \{-1, 1\}$,

$$\frac{1}{2}\int_a^b U_{\nu,1} v = \frac{\alpha}{2}\int_a^b v = \frac{\alpha}{2}\int_a^b (\gamma^2 v'' + \alpha) = \frac{\alpha\gamma^2}{2}[v'(b) - v'(a)] + \frac{\alpha^2}{2}(b - a).$$

This implies

$$\begin{aligned}\mathcal{L}(U_{\nu,1}) &= -\gamma^2[v'(x_1) - v'(x_2) + v'(x_3) - \cdots + (-1)^{\nu+1} v'(x_\nu)]\\ &= -\gamma^2 \nu v'(x_1).\end{aligned}$$

We now need to calculate $v'(x_1)$. On $(0, x_1)$ $v(x) = -1 + C' \cosh(x/\gamma)$ and on (x_1, x_2) $v(x) = 1 + C'' \cosh((x - \frac{x_1+x_2}{2})/\gamma)$ for some appropriate C' and C''. They and their derivatives match at x_1. Therefore,

$$\begin{cases} -1 + C' \cosh\left(\dfrac{x_1}{\gamma}\right) = 1 + C'' \cosh\left(\dfrac{x_1 - x_2}{2\gamma}\right), \\[2ex] \dfrac{C'}{\gamma}\sinh\left(\dfrac{x_1}{\gamma}\right) = \dfrac{C''}{\gamma}\sinh\left(\dfrac{x_1 - x_2}{2\gamma}\right). \end{cases}$$

Solving this system we find

$$\begin{cases} C' = \dfrac{2\sinh(\frac{x_2-x_1}{2\gamma})}{\sinh(\frac{1}{\gamma\nu})}, \\[2ex] C'' = \dfrac{-2\sinh(\frac{x_1}{\gamma})}{\sinh(\frac{1}{\gamma\nu})}. \end{cases}$$

Recall that $x_1 = \frac{1-m}{2\nu}$, $x_2 - x_1 = \frac{1+m}{\nu}$, and $\frac{x_1+x_2}{2} = \frac{1}{\nu}$. Then

$$v'(x_1) = \frac{2\sinh(\frac{1+m}{2\gamma\nu})\sinh(\frac{1-m}{2\gamma\nu})}{\gamma\sinh(\frac{1}{\gamma\nu})}.$$

Going back to $\mathcal{L}(U_{\nu,1})$, we find

$$\mathcal{L}(U_{\nu,1}) = \mathcal{L}(U_{\nu,2}) = -\frac{2\gamma\nu\sinh(\frac{1+m}{2\gamma\nu})\sinh(\frac{1-m}{2\gamma\nu})}{\sinh(\frac{1}{\gamma\nu})},$$

and hence the key formula

$$\mathcal{J}_0(U_{\nu,1}) = \mathcal{J}_0(U_{\nu,2}) = c_0\nu - \frac{2\gamma\nu\sinh(\frac{1-m}{2\gamma\nu})\sinh(\frac{1+m}{2\gamma\nu})}{\sinh(\frac{1}{\gamma\nu})}, \tag{4.3}$$

analogous to (3.9).

We need to show that (1) $\mathcal{L}(U_{\nu,1}) = \mathcal{L}(U_{\nu,2})$ is decreasing in ν in order to prove Proposition 3.1 for $\mathcal{J}_0$, and (2) $\mathcal{J}_0(U_{\nu,1}) = \mathcal{J}_0(U_{\nu,2})$ is convex in ν in order to prove Theorem 1.2 for $\mathcal{J}_\epsilon$.

For $t > 0$ define

$$g(t) = c_0 t - \frac{2\gamma t\sinh(\frac{1-m}{2\gamma t})\sinh(\frac{1+m}{2\gamma t})}{\sinh(\frac{1}{\gamma t})}.$$

Compute g' and g''.

$$\begin{aligned} g'(t) = c_0 - \frac{2\gamma}{t(\sinh\frac{p_1+p_2}{t})^2}\Big\{ & t\sinh\frac{p_1}{t}\sinh\frac{p_2}{t}\sinh\frac{p_1+p_2}{t} \\ & -p_1\cosh\frac{p_1}{t}\sinh\frac{p_2}{t}\sinh\frac{p_1+p_2}{t} \\ & -p_2\sinh\frac{p_1}{t}\cosh\frac{p_2}{t}\sinh\frac{p_1+p_2}{t} \\ & +(p_1+p_2)\sinh\frac{p_1}{t}\sinh\frac{p_2}{t}\cosh\frac{p_1+p_2}{t}\Big\}, \\ g''(t) = \frac{4\gamma}{(t\sinh\frac{1}{\gamma t})^3}\Big\{ & \cosh\frac{1}{\gamma t}\left[p_1^2\left(\sinh\frac{p_2}{t}\right)^2 + p_2^2\left(\sinh\frac{p_1}{t}\right)^2\right] \\ & -2p_1p_2\sinh\frac{p_1}{t}\sinh\frac{p_2}{t}\Big\}, \end{aligned}$$

where $p_1 = \frac{1-m}{2\gamma}$ and $p_2 = \frac{1+m}{2\gamma}$. Since $\cosh\frac{1}{\gamma t} > 1$,

$$g''(t) > \frac{4\gamma}{(t\sinh\frac{1}{\gamma t})^3}\left\{p_1^2\left(\sinh\frac{p_2}{t}\right)^2 + p_2^2\left(\sinh\frac{p_1}{t}\right)^2 - 2p_1p_2\sinh\frac{p_1}{t}\sinh\frac{p_2}{t}\right\} \geq 0.$$

Like (3.9), $g(t)$ is convex in t. So $\mathcal{J}_0(U_{\nu,1}) = \mathcal{J}_0(U_{\nu,2})$ is convex in ν.

To show that $\mathcal{L}(U_{\nu.1}) = \mathcal{L}(U_{\nu,2})$ is decreasing in ν, we set

$$h(t) = -\frac{2\gamma t \sinh(\frac{1-m}{2\gamma t}) \sinh(\frac{1+m}{2\gamma t})}{\sinh(\frac{1}{\gamma t})}.$$

Then $h' = g' - c_0$ and $h'' = g''$. Near $t = \infty$ we can find Taylor's expansion

$$h'(t) = -\frac{(1-m^2)^2}{12\gamma^2}\frac{1}{t^3} + o\left(\frac{1}{t^3}\right).$$

Therefore $\lim_{t\to\infty} h'(t) = 0$. Then $h'' > 0$ implies that $h'(t) < 0$ for all $t > 0$. So $\mathcal{L}(U_{\nu.1}) = \mathcal{L}(U_{\nu,2})$ is decreasing in ν. We also note

$$\lim_{\nu\to\infty} \mathcal{L}(U_{\nu,1}) = \lim_{\nu\to\infty} \mathcal{L}(U_{\nu,2}) = \lim_{t\to\infty} h(t) = -\frac{1-m^2}{2}.$$

The global minimum t_0 of $g(t)$, $t \geq 1$, in the case of $\mathcal{J}_0$, cannot be obtained explicitly. But its existence and uniqueness are guaranteed by the convexity of g and the fact $\lim_{t\to\infty} g'(t) = c_0 > 0$. Taylor's expansion

$$g'(t) = c_0 - \frac{(1-m^2)^2}{12\gamma^2}\frac{1}{t^3} + o\left(\frac{1}{t^3}\right)$$

implies that if c_0 is small, then t_0, and ν_* in Theorem 1.2, is close to $(\frac{(1-m^2)^2}{12\gamma^2 c_0})^{1/3}$. This number is the same as the one for $\mathcal{I}_0$.

Therefore it can be argued heuristically that as c_0 becomes small, the problems $\mathcal{I}_\epsilon$ and $\mathcal{J}_\epsilon$ start to converge.

Acknowledgment. Part of the work was done during an RiP stay at Mathematisches Forschungsinstitut Oberwolfach, which the authors would like to thank for its hospitality.

REFERENCES

[1] M. Bahiana and Y. Oono, *Cell dynamical system approach to block copolymers*, Phys. Rev. A (3), 41 (1990), pp. 6763–6771.

[2] J. Carr, M. E. Gurtin, and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Rational Mech. Anal., 86 (1984), pp. 317–352.

[3] G. Dal Maso, *An Introduction to Γ-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser, Boston, MA, 1993.

[4] L. Evans and R. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, New York, London, Tokyo, 1992.

[5] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, New York, Heidelberg, Berlin, 1965.

[6] R. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proc. Royal Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[7] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 357–383.

[8] L. Modica and S. Mortola, *Un esempio di Γ^--convergenza,* Boll. Un. Mat. Ital. B, (5) 14 (1977), pp. 285–299.

[9] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.

[10] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the micro-phase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.

[11] T. Ohta and K. Kawasaki, *Equilibrium morphology of block polymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.

[12] X. Ren and L. Truskinovsky, *Finite Scale Microstructures in Nonlocal Elasticity*, preprint.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 925–939

ON THE LARGE TIME BEHAVIOR OF SOLUTIONS OF HAMILTON–JACOBI EQUATIONS[*]

G. BARLES[†] AND PANAGIOTIS E. SOUGANIDIS[‡]

Abstract. We consider the long time behavior of viscosity solutions of first-order Hamilton–Jacobi equations with periodic space dependence. We prove, under sharp conditions, that as time goes to infinity, solutions converge to solutions of the corresponding stationary equation.

Key words. Hamilton–Jacobi equations, periodicity, ergodic problem, long time behavior, viscosity solutions, hamiltonian systems

AMS subject classifications. 70H20, 58F11, 58F05, 49L25

PII. S0036141099350869

1. Introduction.

In this article we are interested in the behavior, as $t \to +\infty$, of the viscosity solutions of first-order Hamilton–Jacobi equations of the form

$$u_t + H(x, Du) = 0 \quad \text{in } \mathbb{R}^N \times (0, +\infty) \ , \tag{1.1}$$

$$u = u_0 \quad \text{on } \mathbb{R}^N \times \{0\} \ , \tag{1.2}$$

where the hamiltonian H, the initial datum u_0, and the solution u are assumed to be real-valued continuous functions and $Du = (\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_N})$ denotes the gradient of u.

Throughout the paper we suppose that both H and u_0 are $\mathbb{Z}^N$-*periodic* in x, i.e., that for all $x, p \in \mathbb{R}^N$ and $z \in \mathbb{Z}^N$,

$$H(x+z, p) = H(x, p) \quad \text{and} \quad u_0(x+z) = u_0(x) \ . \tag{1.3}$$

We also assume that a comparison (uniqueness) result holds for (1.1)–(1.2). The first consequence of this assumption is the $\mathbb{Z}^N$-periodicity in x of the solution for any $t > 0$.

The study of the long time behavior of solutions of (1.1) and (1.2) first leads to an *ergodic* problem. Indeed, the first step is to show the existence of a constant c_0, depending only on H and not u_0, such that the function $u(\cdot, t) + c_0 t$ remains bounded, as $t \to +\infty$. The classical result in this direction is due to Lions, Papanicolaou, and Varadhan [7], who obtained the existence of such a constant c_0, the so-called *ergodic cost*, under the following coercivity assumption on H:

$$H(x, p) \to +\infty \quad \text{when } |p| \to +\infty, \text{ uniformly in } x \in \mathbb{R}^N. \tag{1.4}$$

Another way to define c_0 is by using the stationary Hamilton–Jacobi equation. Indeed, c_0 is the unique constant c for which the equation

$$H(x, Du) = c \quad \text{in } \mathbb{R}^N \tag{1.5}$$

[*] Received by the editors February 2, 1999; accepted for publication July 28, 1999; published electronically April 4, 2000. This work was partially supported by the European TMR program "Viscosity Solutions and Their Applications."

http://www.siam.org/journals/sima/31-4/35086.html

[†] Laboratoire de Mathématiques et Physique Théorique, Faculté des Sciences et Techniques Université de Tours, Parc de Grandmont, 37200 Tours, France (barles@univ-tours.fr).

[‡] Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 (souganid@math.wisc.edu). This author was partially supported by the NSF, ONR, and the TMR programs on "Viscosity Solutions and Their Applications" and "Hyperbolic Conservation Laws."

has a continuous, periodic viscosity solution.

It is also worth pointing out that if $u_0 \in W^{1,\infty}(\mathbb{R}^N)$, then (1.4) yields that the function $(x,t) \mapsto u(x,t) + c_0 t$ is in $W^{1,\infty}(\mathbb{R}^N \times [0,+\infty))$. This is a key fact, since it provides the compactness in $C(\mathbb{R}^N)$ of the functions $u(\cdot,t) + c_0 t$ for $t > 0$, a fact which is essential in the study of the behavior of these functions when $t \to +\infty$. Before coming back to this question, which is the central purpose of our work, we mention that such an ergodic problem in the deterministic control framework was systematically studied by Arisawa [1, 2].

In this article we are interested in the next step, i.e., in the behavior of $u(\cdot,t)+c_0 t$ as $t \to +\infty$. To simplify the exposition, we are going to assume, without any loss of generality, that $c_0 = 0$. With this convention, the question we address here can be formulated in the following way:

Assume that $u \in W^{1,\infty}(\mathbb{R}^N \times (0,+\infty))$ or $u \in BUC(\mathbb{R}^N \times (0,+\infty))$. Is it true that, as $t \to \infty$,

$$u(\cdot,t) \to u_\infty(\cdot) \quad \textit{in } C(\mathbb{R}^N)\ , \tag{1.6}$$

where u_∞ is a viscosity solution of the stationary equation

$$H(x,Du) = 0 \quad \textit{in } \mathbb{R}^N\ ? \tag{1.7}$$

The apparent simplicity of this question is misleading. In fact, this problem has remained open for a long time. The first results on the asymptotic behavior of viscosity solutions of Hamilton–Jacobi equations were obtained in Lions [6] and in Barles [3] essentially for either x-independent cases or equations involving a suitable dependence on u. It was only very recently that Namah and Roquejoffre [8] and Fathi [5] succeeded in proving rather general results related to the above questions, which we now briefly describe and compare.

The results of [5] and [8] are obtained for equations set on compact manifolds and for hamiltonians which are convex in the p-variable and satisfy (1.4). The result of [5] was proved under the additional assumption that H is smooth and strictly convex; i.e., there exists a constant $\alpha > 0$ such that

$$D^2_{pp}H(x,p) \geq \alpha I \quad \text{ in } \mathbb{R}^N \times \mathbb{R}^N\ . \tag{1.8}$$

The proof relies on the representation of the solution u by the so-called Oleinik–Lax formula and is based on dynamical systems methods. In particular, [5] emphasizes the central role played by the Aubry–Mather set, an attractor set for the geodesics associated with the Lax–Oleinik formula. This result was revisited recently by Roquejoffre [10], who uses a combination of partial differential equations and dynamical systems methods. (See also Roquejoffre [9] for results in dimension 1.)

The approach of [8] is based on partial differential equations methods and requires a condition, which in the $\mathbb{R}^N$-framework can be stated as follows:

$$\begin{cases} \text{There exists a } C^1\text{-function } \phi \in BUC(\mathbb{R}^N) \text{ such that} \\ H(x,D\phi(x)) \leq 0 \quad \text{ in } \mathbb{R}^N \text{ and} \\ \qquad H(x,p+D\phi(x)) > H(x,D\phi(x)) \text{ for all } x \in \mathbb{R}^N \text{ and } p \in \mathbb{R}^N\backslash\{0\}. \end{cases} \tag{1.9}$$

The two key arguments of [8] are that $u(\cdot,t)$ is decreasing (and therefore uniformly convergent) on the set $K = \{x \in \mathbb{R}^N : H(x,D\phi(x)) = 0\}$—note that this set is necessarily a nonempty subset of $\mathbb{R}^N$ as a consequence of the fact that $c_0 = 0$—and that there is a strong comparison principle for the Dirichlet problem

$$H(x, Dw) = 0 \text{ in } \mathbb{R}^N \backslash K, \qquad w = \varphi \text{ on } \partial K,$$

where φ is a continuous function. This last property holds because ϕ is a (local) strict subsolution of the equation in $\mathbb{R}^N \backslash K$. The strong comparison principle for viscosity solutions then allows the use of the half-relaxed limits methods.

Here we provide a generalization of these two types of results. In particular, we are able to treat hamiltonians which are not necessarily convex and to remove the assumptions on the regularity of H and ϕ. Moreover the $\mathbb{Z}^N$-periodic setting we chose here for the sake of simplicity can be replaced without any additional difficulty by a general compact manifold one. It is, however, worth mentioning that some kind of compactness assumption on the domain is necessary at least to apply our strategy of proof.

Our main argument, which is completely different from those given in [5], [8], [9], and [10] can be described roughly in the following way: We first show that

$$||(u_t)^-(\cdot, t)||_{L^\infty(\mathbb{R}^N)} \to 0 \quad \text{as } t \to +\infty \ . \tag{1.10}$$

For the reader's convenience we provide in section 3, under simplified assumptions, a formal argument which shows why such property should be true. In fact, the formulation of the precise results (Theorems 3.1 and 3.2) is a bit more general but unfortunately rather technical. The main consequence of (1.10) is that the ω-limit set of the function $u(\cdot, t)$ contains only subsolutions of (1.7). In turn, this property is enough to prove (1.6). It is in this last step that the compactness property of the domain seems to play a key role.

This paper is organized as follows: In section 2 we state the assumptions and the main results of the paper. Section 3 is devoted to the statement of the weak versions of (1.10) which are proved in the appendix. In section 4 we prove the main results and in section 5 we discuss the main assumptions on the hamiltonian and some extensions.

2. The main results and their applications. To formulate the main results we recall that we are interested in the asymptotic behavior of solutions $u \in BUC(\mathbb{R}^N \times [0, \infty))$ of the initial value problem

$$\begin{cases} u_t + H(x, Du) = 0 & \text{in } \mathbb{R}^N \times (0, \infty), \\ u = u_0 & \text{on } \mathbb{R}^N \times \{0\}, \end{cases} \tag{2.1}$$

under the assumptions that

$$\text{(H1)} \quad \begin{cases} H \text{ is continuous in } \mathbb{R}^N \times \mathbb{R}^N \text{ and } \mathbb{Z}^N\text{-}\textit{periodic} \text{ with respect to } x, \\ \text{i.e., for all } x, p \in \mathbb{R}^N \text{ and } z \in \mathbb{Z}^N, \\ \qquad H(x + z, p) = H(x, p). \end{cases}$$

We also assume the following.

(H2) There exists a viscosity subsolution $\phi \in BUC(\mathbb{R}^N)$ of $H(x, D\phi) \leq 0$ in $\mathbb{R}^N$.

$$\text{(H3)} \quad \begin{cases} \text{Either } u \text{ and } \phi \text{ are in } W^{1,\infty}(\mathbb{R}^N \times (0, \infty)) \text{ or there exists a} \\ \text{continuous function } m : [0, +\infty) \to \mathbb{R}^+ \text{ such that } m(0^+) = 0 \\ \text{and, for all } x, y \in \mathbb{R}^N \text{ and } p \in \mathbb{R}^N, \\ \qquad |H(x, p) - H(y, p)| \leq m(|x - y|(1 + |p|)), \end{cases}$$

and

$$\text{(H4)}\quad \begin{cases} \text{there exist } \eta > 0 \text{ and } \psi(\eta) > 0 \text{ such that, if } H(x, p+q) \geq \eta \text{ and} \\ H(x,q) \leq 0 \text{ for some } x \in A \subset \mathbb{R}^N,\ p, q \in \mathbb{R}^N, \text{ then, for all } \mu \in (0,1], \\ \\ \qquad\qquad \mu H\Big(x, \mu^{-1}p + q\Big) \geq H(x, p+q) + \psi(\eta)(1-\mu). \end{cases}$$

Note that if H is C^1 in p, then (H4) reduces to

$$\text{(H4)}'\quad \begin{cases} \qquad\qquad H_p(x, p+q)\cdot p - H(x,p+q) \geq \psi(\eta), \\ \\ \text{for any } x \in A,\ p, q \in \mathbb{R}^N \text{ such that } H(x,p+q) \geq \eta \text{ and } H(x,q) \leq 0. \end{cases}$$

The final assumption is as follows.

$$\text{(H5)}\quad \begin{cases} \text{There exists a, possibly empty, compact subset } K \text{ of } \mathbb{R}^N \text{ such that} \\ \\ \quad \text{(i)}\quad H(x,p) \geq 0 \text{ on } K \times \mathbb{R}^N, \text{ and} \\ \\ \quad \text{(ii)}\quad \text{for all } \delta > 0, \text{ (H4) holds with } A = (K_\delta)^c \text{ for all } \eta > 0, \\ \qquad\quad \text{where } (K_\delta)^c = \{x \in \mathbb{R}^N : d(x,K) > \delta\} \text{ with } \psi \text{ depending on } \delta. \end{cases}$$

The result about the asymptotic behavior of the solution u of (1.1) and (1.2) is the following.

THEOREM 2.1. *Assume that* (H1)–(H3) *and* (H5) *hold. If* $u \in BUC(\mathbb{R}^N \times (0,\infty))$ *is a* $\mathbb{Z}^N$*-periodic in* x *solution of* (1.1), *then there exists a* $\mathbb{Z}^N$*-periodic* $\overline{u} \in BUC(\mathbb{R}^N)$ *such that*

(i) $H(x, D\bar{u}) = 0$ *in* $\mathbb{R}^N$, *and*

(ii) $u(x,t) \to \bar{u}(x)$, *uniformly in* $\mathbb{R}^N$, *as* $t \to \infty$.

Before stating a variant of this result, which holds under simpler hypotheses on H, we want to point out that it is generally rather difficult to show that the solution u of (1.1)–(1.2) is actually in $BUC(\mathbb{R}^N \times (0,\infty))$. The classical existence results provide a solution which is only in $BUC(\mathbb{R}^N \times (0,T))$ for all $T > 0$ (see Barles [4]). To the best of our knowledge, as we already mentioned in the introduction, the only general result which gives the compactness in $C(\mathbb{R}^N)$ of the functions $u(\cdot, t)$—the important information—is the one obtained by Lions, Papanicolaou, and Varadhan [7] (see also Namah and Roquejoffre [8]) under the following assumption:

(H6) $\qquad H(x,p) \to \infty$ as $|p| \to \infty$, uniformly in $x \in \mathbb{R}^N$.

As a consequence of this, the reader can replace in any of our results the assumption "$u \in W^{1,\infty}(\mathbb{R}^N \times (0,\infty))$" with "(H6) and $u_0 \in W^{1,\infty}(\mathbb{R}^N)$" and in the same way "$u \in BUC(\mathbb{R}^N \times (0,\infty))$" with "(H6) and $u_0 \in BUC(\mathbb{R}^N)$," which implies in both cases the existence of such a solution.

It is also worth mentioning that if we assume that the sets $\{p \in \mathbb{R}^N : H(x,p) \leq 0\}$ are bounded uniformly for $x \in \mathbb{R}^N$ and that (H4) holds with $A = \mathbb{R}^N$, then (H6) are a direct consequence of (H4). Hence, in these cases we do not lose any generality by assuming (H6).

Finally we write about the existence of a BUC-subsolution ϕ as an assumption. In fact, this property together with the global boundedness of u is a direct consequence of the definition of c_0 (recall that we assume $c_0 = 0$).

To state a variant of Theorem 2.1 we introduce the following simpler hypotheses.

$$\text{(H7)}\quad \begin{cases} \text{There exists a family } (\phi_\varepsilon)_{\varepsilon>0} \text{ of } C^1(\mathbb{R}^N) \cap W^{1,\infty}(\mathbb{R}^N)\text{-functions, which} \\ \text{are uniformly bounded in } \varepsilon > 0 \text{ and satisfy} \\ \\ \qquad\qquad H(x, D\phi_\varepsilon) \leq \varepsilon \quad \text{in } \mathbb{R}^N\ . \end{cases}$$

(H8) For every $x \in \mathbb{R}^N$, the function $p \mapsto H(x,p)$ is locally Lipschitz.

$$\text{(H9)}\quad\begin{cases}\text{There exists a, possibly empty, compact subset } K \text{ of } \mathbb{R}^N \text{ such that}\\ \text{(i)}\ \ H(x,p)\geq 0 \text{ on } K\times\mathbb{R}^N, \text{ and}\\ \text{(ii)}\ \text{if } H(x,p)\geq\eta>0 \text{ and } d(x,K)\geq\eta, \text{ then for all sufficiently}\\ \qquad \text{small, compared to } \eta,\ \varepsilon>0,\\ H_p(x,p)\cdot(p-D\phi_\varepsilon(x))-H(x,p)\geq\psi(\eta)>0 \ \text{ for all } x \text{ and a.e. in } p.\end{cases}$$

We have the following theorem.

THEOREM 2.2. *Assume* (H1) *and* (H7)–(H9). *Then any $\mathbb{Z}^N$-periodic in x solution $u\in W^{1,\infty}(\mathbb{R}^N\times(0,\infty))$ of* (2.1) *converges, uniformly in x, as $t\to\infty$, to a $\mathbb{Z}^N$-periodic in x solution $\bar{u}$ of* (1.7).

The differences between Theorem 2.1 and Theorem 2.2 (which are not so obvious at first glance) will become clear in their proofs. Indeed, to prove Theorem 2.2 we will use Theorem 3.2, the proof of which is far simpler than the one of Theorem 3.1, which provides the key argument in the proof of Theorem 2.1.

On the other hand, for hamiltonians H, which are convex in p, and for Lipschitz continuous solutions, Theorem 2.2 is as general as Theorem 2.1 since, in particular, the existence of the ϕ_ε can be obtained from (H2) by a standard regularization argument.

Here we discuss in detail the two classes of examples presented in the introduction and show how they follow from the above theorems. We also present an example not covered by [5], [8], [9], and [10].

(i) *The Namah–Roquejoffre case.* The hamiltonian H is assumed to be of the form

$$H(x,p)=F(x,p)-f(x), \tag{2.2}$$

where

$$\begin{cases}F\in C(\mathbb{R}^N\times\mathbb{R}^N) \text{ is } \mathbb{Z}^N\text{-periodic in } x, \text{ convex in } p, \text{ and}\\ F(x,p)>F(x,0)\equiv 0 \text{ for all } x\in\mathbb{R}^N \text{ and } p\in\mathbb{R}^N\backslash\{0\},\end{cases} \tag{2.3}$$

and

$$\begin{cases}f \text{ is continuous, } \mathbb{Z}^N\text{-periodic in } x,\ f\geq 0 \text{ on } \mathbb{R}^N, \text{ and}\\ K_f=\{x\in\mathbb{R}^N : f(x)=0\} \text{ is a nonempty compact subset of } \mathbb{R}^N.\end{cases} \tag{2.4}$$

In [8] the authors considered Lipschitz continuous solutions. Hence (H1), (H3) are clearly satisfied, while (H7) is satisfied with $\phi_\varepsilon\equiv 0$. It only remains to check (H8) and (H9). To this end, we observe that the convexity of H in p yields (H8). Moreover, for all x and for almost all $p\in\mathbb{R}^N$,

$$H(x,0)\geq H(x,p)+H_p(x,p)(0-p),$$

and, hence,

$$H_p(x,p)\cdot p-H(x,p)\geq H(x,0)=f(x).$$

It follows that (H9) is satisfied with $K=K_f$. Indeed, it is clear that if $d(x,K_f)\geq\delta$, then $f(x)\geq\psi(\delta)>0$, with ψ independent of η.

Finally it is worth pointing out that the result is also true when (H7) holds with $\phi_\varepsilon\equiv\phi\in C^1(\mathbb{R}^N)\cap W^{1,\infty}(\mathbb{R}^N)$. This is a consequence of the above analysis after changing u to $u-\phi$ in the equation.

(ii) *The Fathi case and extensions.* The main assumption on H in this case is that

$$\left\{ \begin{array}{l} H \text{ is } C^1(\mathbb{R}^N \times \mathbb{R}^N) \text{ and there exists } \alpha > 0 \text{ such that} \\ \qquad\qquad D^2_{pp}H(x,p) \geq \alpha \text{ Id} \text{ in } \mathcal{D}'(\mathbb{R}^N \times \mathbb{R}^N). \end{array} \right. \tag{2.5}$$

It is immediate from (2.5), at least when $H \in C^2(\mathbb{R}^N \times \mathbb{R}^N)$, that for all $x, q \in \mathbb{R}^N$ and for almost all $p \in \mathbb{R}^N$,

$$H(x,q) \geq H(x,p) + H_p(x,p) \cdot (q-p) + \frac{\alpha}{2}|p-q|^2,$$

and, hence,

$$H_p(x,p+q) \cdot p - H(x,p+q) \geq -H(x,q) + \frac{\alpha}{2}|p|^2.$$

If $H(x,q) \leq 0$ and $H(x,p+q) \geq \eta$, it is clear that there exists $\psi(\eta) > 0$ such that $|p| \geq \psi(\eta)$. In this case we may take $K = \emptyset$ in (H9). When H is not smooth, we argue by approximations. Note that in [5] H is assumed to be smooth.

(iii) *Another example.* Consider a hamiltonian H of the form

$$H(x,p) = \psi(x,p)F\Big(x, \frac{p}{|p|}\Big) - f(x), \tag{2.6}$$

where $f \in C(\mathbb{R}^N)$ is nonnegative and $\mathbb{Z}^N$-periodic in x; $F \in C(\mathbb{R}^N \times \mathbb{R}^N \backslash \{0\})$ is continuous, strictly positive, bounded, and $\mathbb{Z}^N$-periodic in x; and $\psi(x,p) = |p + q(x)|^2 - |q(x)|^2$, where $q \in C(\mathbb{R}^N)$ is $\mathbb{Z}^N$-periodic in x. In addition we assume that for some $x_0 \in \mathbb{R}^N$, $q(x_0) = 0$ and $f(x_0) = 0$. It turns out then that $c_0 = 0$ and ϕ can be chosen to be any constant.

It is a bit tedious but straightforward to check that H satisfies the assumptions of Theorem 2.1 but neither (2.3)–(2.4) nor (2.5).

3. Some preliminary results.

Here we present two results about the behavior in time, and for large times, of solutions of (1.1). These results are of independent interest themselves. Their proofs, however, are rather technical. In order not to confuse the issue here and for the reader's convenience, we present them in the appendix.

Both results hold for hamiltonians which are not necessarily periodic in x. Instead of restating the assumption of the previous section here, without the (H1) and for any domain, we first introduce the assumption that

(H1)′ H is uniformly continuous on $\mathbb{R}^N \times \overline{B}_R$, for all $R > 0$ where $\overline{B}_R = \{p \in \mathbb{R}^N : |p| \leq R\}$

and summarize the other hypotheses as follows:

(H10) $\left\{ \begin{array}{l} \text{(H1)}', \text{(H2), (H3), and (H4) hold for } (x,p) \in \Omega \times \mathbb{R}^N, \text{ with } \Omega \text{ a given open} \\ \text{subset of } \mathbb{R}^N, \text{ and } w \in BUC(\bar{\Omega} \times [0,\infty)) \text{ a solution of (1.1) in } \Omega \times (0,\infty). \end{array} \right.$

Before we state the main result, we remark that we may assume, without any loss of generality, that

$$w - \phi \geq 1 \quad \text{in } \bar{\Omega} \times [0, +\infty). \tag{3.1}$$

Indeed, the form of H allows the change ϕ to $\phi - K$ for any constant K. Since $w \in BUC(\bar{\Omega} \times [0,\infty))$, to achieve (3.1) it suffices to choose K sufficiently large.

We also need to introduce, for $\eta > 0$, the functions

$$\mu_\eta(t) = \min_{x\in\bar{\Omega}, s\geq t}\Big[\frac{w(x,s)-\phi(x)+2\eta(s-t)}{w(x,t)-\phi(x)}\Big], \tag{3.2}$$

and

$$\chi_\eta(t) = \min_{x\in\partial\Omega, s\geq t}\Big[\frac{w(x,s)-\phi(x)+2\eta(s-t)}{w(x,t)-\phi(x)}\Big]. \tag{3.3}$$

It follows easily that μ_η, $\chi_\eta : [0,\infty) \to \mathbb{R}$ are uniformly continuous and that $0 \leq \mu_\eta \leq \chi_\eta \leq 1$. Finally, if $\Omega = \mathbb{R}^N$, we define $\chi_\eta \equiv -\infty$.

The first result is the following.

Theorem 3.1. *Assume* (H10). *Then there exists a constant C depending only on w and ϕ such that*

$$\mu_\eta(t) \geq 1 + \inf_{\theta\leq t}[(\chi_\eta(\theta)-1)e^{-C\psi(\eta)(t-\theta)}] \wedge (\mu_\eta(0)-1)e^{-C\psi(\eta)t}. \tag{3.4}$$

Moreover, if $w\big|_{\partial\Omega}$ converges, uniformly in $x \in \partial\Omega$, as $t \to \infty$, then, for all $s \geq t$ and $x \in \bar{\Omega}$,

$$w(x,t) - w(x,s) - 2\eta(s-t) \leq \delta_\eta(t),$$

where $\delta_\eta : [0,\infty) \to [0,\infty)$ is such that $\delta_\eta(t) \to 0$ as $t \to \infty$.

We continue with some preliminaries for the second result, which is also proved in the appendix. To this end, we consider the solution $w \in BUC(\overline{\Omega} \times (0,\infty))$ of the equation

$$\frac{\partial w}{\partial t} + F(x,w,Dw) = 0 \quad \text{in } \Omega \times (0,\infty), \tag{3.5}$$

where

(H11) $\qquad F$ is uniformly continuous on $\bar{\Omega} \times [-R,R] \times \bar{B}_R$, for all $R > 0$,

$$\text{(H12)}\quad \begin{cases} \text{either } w \in W^{1,\infty}(\mathbb{R}^N \times (0,\infty)) \text{ or for each } R > 0, \text{ there exists a} \\ \text{continuous function } m_R : [0,\infty) \to [0,\infty) \text{ such that } m_R(0^+) = 0 \\ \text{and, for all } x, y \in \bar{\Omega},\ p \in \mathbb{R}^N, \text{ and } w \in [-R,R], \\ \qquad |F(x,w,p) - F(y,w,p)| \leq m_R(|x-y|(1+|p|)), \end{cases}$$

and

$$\text{(H13)}\qquad \frac{\partial F}{\partial w}(x,w,p) \geq \psi(\eta) > 0 \text{ a.e., if } F(x,w,p) \geq \eta > 0.$$

To state the result we need to introduce, for $\eta > 0$, the functions

$$M_\eta(t) = \sup_{x\in\bar{\Omega}, s\geq t}[w(x,t) - w(x,s) - 2\eta(s-t)] \tag{3.6}$$

and

$$X_\eta(t) = \sup_{x\in\partial\Omega, s\geq t}[w(x,t) - w(x,s) - 2\eta(s-t)], \tag{3.7}$$

with the convention that $X_\eta = -\infty$ if $\partial\Omega = \phi$.

We have the following theorem.

THEOREM 3.2. *Assume* (H11), (H12), *and* (H13). *Then for all* $\eta > 0$ *and* $t \geq 0$,

$$M_\eta(t) = \sup_{\theta \leq t}[X_\eta(\theta)e^{-\psi(\eta)(t-\theta)}] \vee M_\eta(0)e^{-\psi(\eta)t}. \tag{3.8}$$

Moreover, if $w\big|_{\partial\Omega}$ *converges, uniformly in* $x \in \partial\Omega$, *as* $t \to \infty$, *then, for all* $s \geq t$ *and* $x \in \bar{\Omega}$, *we have*

$$w(x,t) - w(x,s) - 2\eta(s-t) \leq \delta_\eta(t),$$

where $\delta_\eta : [0,\infty) \to [0,\infty)$ *is such that* $\delta_\eta(t) \to 0$ *as* $t \to \infty$.

The conclusions of Theorems 3.1 and 3.2 are in some sense weak versions of (1.10). For the reader's convenience we present below a formal argument, which explains why (1.10) should hold.

To this end let us assume that w is a smooth solution of (3.5) in $\mathbb{R}^N \times (0,\infty)$. A straightforward application of the maximum principle yields that the function $t \mapsto \|(w_t)^-\|_\infty$ is decreasing in time. If (1.10) were not true, then there must exist some $\eta > 0$ and t_0 such that for all $t \geq t_0$,

$$\|(w_t)^-\|_\infty \geq \eta. \tag{3.9}$$

Let $z = w_t$ and $m(t) = \|z^-\|_\infty$. Differentiating (3.5) with respect to t, we find that

$$z_t + F_w(x,w,t,Dw)z + D_pF \cdot Dz = 0.$$

It then follows that

$$m' + F_w(x,w,t,Dw)m = 0.$$

Using (3.9) and (H13) we find

$$m' + \psi(\eta)m = 0,$$

which yields

$$m(t) = m(t_0)e^{-\psi(\eta)(t-t_0)}.$$

Letting $t \to \infty$ contradicts (3.9).

4. The proofs of Theorems 2.1 and 2.2. We begin with the following proof.

Proof of Theorem 2.1. 1. Assumption (H5) yields that u is decreasing on K and, hence, $u\big|_K$ converges, uniformly in x, as $t \to \infty$.

2. Consider the function χ_η defined by (3.3), with $\Omega = (K_\eta)^c$, where, for $\eta > 0$, $K_\eta = \{x \in \mathbb{R}^N : d(x,K) \geq \eta\}$. Step 1 and the uniform of continuity of u then imply that

$$\varliminf_{t\to\infty} \chi_\eta(t) \geq 1 - \nu(\eta), \quad \text{where} \quad \nu(\eta) \to 0 \quad \text{as} \quad \eta \to 0.$$

Using this last observation and applying Theorem 3.1 we obtain that

$$\varliminf_{t\to\infty} \mu_\eta(t) \geq 1 - \tilde{\nu}(\eta), \quad \text{where} \quad \tilde{\nu}(\eta) \to 0 \quad \text{as} \quad \eta \to 0.$$

Therefore, for all $s > 0$ and for all $(x,t) \in (K_\eta)^c \times [0,s]$, we have

$$u(x,t) - u(x,s) - 2\eta(s-t) \le \tilde{\delta}_\eta(t), \tag{4.1}$$

where

$$\overline{\lim_{t\to\infty}}\, \tilde{\delta}_\eta(t) \le \tilde{\nu}(\eta).$$

3. Since the family $(u(\cdot,t))_{t\ge 0}$ is compact in $BUC(\mathbb{R}^N)$ and the functions $u(\cdot,t)$ are periodic in x for all t, we may consider a subsequence $u(\cdot,T_n)$, with $T_n \to +\infty$, converging uniformly in $\mathbb{R}^N$.

The maximum principle for viscosity solutions implies that for any $n,p \in \mathbb{N}$, we have

$$\|u(\cdot,T_n+\cdot) - u(\cdot,T_p+\cdot)\|_{L^\infty(\mathbb{R}^N\times(0,\infty))} \le \|u(\cdot,T_n) - u(\cdot,T_p)\|_{L^\infty(\mathbb{R}^N)}\ . \tag{4.2}$$

It follows from this inequality that $(u(\cdot,T_n+\cdot))_n$ is a Cauchy sequence in $BUC(\mathbb{R}^N\times(0,\infty))$ and therefore it converges uniformly to a function $u_\infty \in BUC(\mathbb{R}^N\times(0,\infty))$.

4. Using (4.1) we find, for all $0 \le t \le s$ and for all $x \in (K_\eta)^c$,

$$u(x,t+T_n) - u(x,s+T_n) - 2\eta(s-t) \le \tilde{\delta}_\eta(t+T_n).$$

Letting $n\to\infty$ and then $\eta \to 0$ yields, for all $0 \le t \le s$ and for all $x \in (K)^c$,

$$u_\infty(x,t) - u_\infty(x,s) \le 0, \tag{4.3}$$

i.e., that u_∞ is increasing in t for $x \in (K)^c$.

On an other hand, step 1 yields that $u\big|_K$ converges, uniformly in x, as $t\to\infty$. Hence u_∞ is constant in time on K.

5. The stability property of viscosity solutions applied to the sequence $(u(\cdot,T_n+\cdot))_n$ then implies that u_∞ is a solution of

$$(u_\infty)_t + H(x,Du_\infty) = 0 \quad \text{in} \quad \mathbb{R}^N\times(0,\infty),$$

and, since u_∞ is increasing in t for all $x\in\mathbb{R}^N$,

$$H(x,Du_\infty(\cdot,t)) \le 0 \quad \text{in } \mathbb{R}^N\times\{t\} \text{ and for all } t>0.$$

Again, the stability implies

$$H(x,Du_\infty(\cdot,0)) \le 0 \quad \text{in } \mathbb{R}^N.$$

This last assertion shows that *any function in the ω-limit set of u is a subsolution of the stationary equation in $\mathbb{R}^N$.*

6. The uniform convergence of $u(\cdot,T_n+\cdot)$ to u_∞ on $\mathbb{R}^N\times(0,\infty)$ yields

$$-o_n(1) + u_\infty(x,t) \le u(x,T_n+t) \le u_\infty(x,t) + o_n(1) \quad \text{in } \mathbb{R}^N. \tag{4.4}$$

Since $u_\infty \in BUC(\mathbb{R}^N\times(0,\infty))$ is increasing with respect to t, it follows that $u_\infty(\cdot,t) \to \overline{u}(\cdot)$, uniformly in x, as $t\to\infty$.

Finally, taking the relaxed half-limits $\limsup^*$ and $\liminf_*$ in t[1] in (4.4) yields

$$-o_n(1) + \overline{u}(x) \le \underline{\lim}_* u(x) \le \overline{\lim}^* u(x) \le \overline{u}(x) + o_n(1) \quad \text{in } \mathbb{R}^N.$$

[1] For $z \in BUC(\mathbb{R}^N\times(0,\infty))$, $\limsup^* z(x) = \limsup_{\substack{y\to x\\ t\to\infty}} z(y,t)$ and $\liminf_* z(x) = \liminf_{\substack{y\to x\\ t\to\infty}} z(y,t)$.

Letting $n \to +\infty$, we obtain

$$\underline{\lim}_* u = \overline{\lim}^* u = \overline{u} \quad \text{in } \mathbb{R}^N,$$

which yields the *uniform convergence of* $u(\cdot,t)$ *to* $\overline{u}(\cdot)$ *as* $t \to \infty$.

7. Finally, by the stability result, the limit of u, as $t \to \infty$, which we still denote by $\overline{u}$, is a (viscosity) solution of $H(x, D\overline{u}) = 0$ in $\mathbb{R}^N$.

We continue with the following proof.

Proof of Theorem 2.2. 1. For each $\epsilon > 0$, we define

$$w^\varepsilon = -\exp[-(u - \phi_\varepsilon)].$$

It is then immediate that

$$w_t^\epsilon + F(x, w^\epsilon, Dw^\epsilon) = 0 \quad \text{in } \mathbb{R}^N \times (0,\infty),$$

where, for $x, p \in \mathbb{R}^N$ and $w \in \mathbb{R}$,

$$F(x,w,p) = -wH\left(x, -\frac{p}{w} + D\phi_\epsilon(x)\right). \tag{4.5}$$

After this change of variable, the proof consists essentially in following readily the proof of Theorem 2.1, replacing only the use of Theorem 3.1 by the use of Theorem 3.2.

2. It is straightforward to verify that F satisfies assumptions (H11), (H12), and (H13) of Theorem 3.2. Therefore, for all $\eta > 0$, for all $x \in (K_\eta)^c$, and for all $t \geq 0$, we have

$$w_\epsilon(x,t) - w_\epsilon(x,s) - \eta(s-t) \leq \tilde{\delta}_\eta(t)\ , \tag{4.6}$$

where $\tilde{\delta}_\eta(t) \to 0$ as $t \to +\infty$.

3. The functions w_ε are uniformly bounded, since u is bounded and the ϕ_ε's are uniformly bounded in ε. Since

$$u(x,t) - u(x,s) = -\log(-w_\epsilon(x,t)) + \log(-w_\epsilon(x,s))\ ,$$

there exists a constant $\tilde{C} > 0$ such that, for all $x \in (K_\eta)^c$ and for all $s \geq t$, we have

$$u(x,t) - u(x,s) \leq \tilde{C}\left[\eta(s-t) + \tilde{\delta}_\eta(t)\right]\ .$$

4. Using this last inequality, the conclusion follows by applying readily the arguments of the proof of Theorem 2.1.

5. Remarks and extensions. A natural question is whether one needs an assumption like, for example, (H4)$'$, on H. The example of the eikonal equation

$$u_t + |Du| = 0 \quad \text{in } \mathbb{R}^N \times (0,+\infty) \tag{5.1}$$

shows that, if we restrict our attention to convex hamiltonians, (H5)(ii) is not necessary to obtain the convergence. Indeed, in this example, we can apply either the result of [8] or Theorem 2.1, since assumption (H5)(i) holds with $K = \mathbb{R}^N$.

The following one-dimensional example shows, however, that, except for equations like (5.1), such an extension does not seem to be possible. Indeed, consider the problem

$$\begin{cases} u_t + |u_x + \alpha| - |\alpha| = 0 & \text{in } \mathbb{R} \times (0,+\infty), \\ u(x,0) = \sin(x) & \text{in } \mathbb{R}. \end{cases} \tag{5.2}$$

If $\alpha > 1$ it is easily checked that the unique viscosity solution of (5.2) is

$$u(x,t) = \sin(x-t) \ ,$$

which is clearly in $W^{1,\infty}(\mathbb{R}^N \times (0,+\infty))$ but does not converge as $t \to +\infty$. For the hamiltonian $H(p) = |p+\alpha| - |\alpha|$, the quantity $H_p \cdot p - H$ vanishes for p such that $\alpha(p+\alpha) \geq 0$ and, therefore, it does not satisfy any of the (H4)-type assumptions.

On the contrary, we remark that Theorem 2.2 applies to the equation

$$u_t + |u_x + \alpha|^2 - |\alpha|^2 = 0 \quad \text{in } \mathbb{R} \times (0,+\infty) \ ,$$

which essentially has the same limiting equation as (5.2), in the sense that both limiting equations have the same viscosity solutions. This example shows that some kind of strict convexity-type property is really playing a role in the asymptotic behavior of the solutions of Hamilton–Jacobi equations.

Typically assumption (H4)$'$ implies that the set $\{p \in \mathbb{R}^N : H(x,p+q) \leq 0\}$ is starshaped. This geometric condition alone does not seem to be sufficient as is shown by (5.2) above. On the other hand, if H is strictly convex and $c_0 = 0$, then any function F which equals H on the set $\{H > 0\}$ and is strictly negative on the set $\{H < 0\}$ satisfies the assumptions of Theorem 2.1.

Appendix. To prove Theorem 3.1 we need the following.

LEMMA A.1. *Under the assumptions of Theorem* 3.1, *the function* μ_η *defined by* (3.2) *is a viscosity solution of the variational inequality*

$$\max(\mu_\eta'(t) + C\psi(\eta)(\mu_\eta(t) - 1), \mu_\eta(t) - \chi_\eta(t)) \geq 0 \quad \textit{in } (0,+\infty). \tag{A.1}$$

Assuming for the moment this lemma we proceed with the following.

Proof of Theorem 3.1. 1. The first part of the claim follows from the facts that (A.1) admits a comparison principle and the right-hand side of (3.4) is a solution of the variational inequality (A.1) with initial datum $\mu_\eta(0)$.

2. The uniform convergence of $w\big|_{\partial\Omega}$ implies that $\chi_\eta(t) \to 1$ as $t \to \infty$. Then (3.4) yields that $\mu_\eta(t) \to 1$ as $t \to \infty$.

It then follows for all $x \in \overline{\Omega}$ and all $s \geq t$ that

$$\mu_\eta(t)(w(x,t) - \phi(x)) - w(x,s) + \phi(x) - 2\eta(s-t) \leq 0,$$

and, hence,

$$w(x,t) - w(x,s) - 2\eta(s-t) \leq \max_{x \in \bar{\Omega}}((1-\mu_\eta(t))(w(x,t) - \phi(x)).$$

It is now clear that the right-hand side of this last inequality tends to 0 as $t \to \infty$. □

Proof of Lemma A.1. 1. Let $\widetilde{\psi} \in C^1((0,+\infty))$ and t be a strict local minimum point of $\mu - \widetilde{\psi}$. Since there is nothing to check if $\mu_\eta(t) \geq \chi_\eta(t)$, we may assume that $\mu_\eta(t) < \chi_\eta(t)$, and, in particular, $\mu_\eta(t) < 1$.

2. For $\varepsilon > 0$ and $\alpha > 0$ we introduce the function

$$\Psi^{\varepsilon,\alpha}(x,y,z,t,s) = \frac{w(x,s) - \phi(z) + 2\eta(s-t)}{w(y,t) - \phi(z)} + \frac{|x-y|^2}{2\varepsilon} + \frac{|x-z|^2}{2\varepsilon} - \widetilde{\psi}(t) + \alpha|x|^2.$$

Classical arguments from the theory of viscosity solutions (see, for example, Barles [3]) yield that the function $\Psi^{\varepsilon,\alpha}$ achieves its minimum over $\overline{\Omega}\times\overline{\Omega}\times\overline{\Omega}\times\{(\tau,s)\backslash s\geq \tau, \tau\in[t-\delta,t+\delta]\}$ at some point $(\bar x,\bar y,\bar z,\bar t,\bar s)$ (as usual we drop the dependence of $\bar x,\bar y,\bar z,\bar t$, and $\bar s$ in ε and α for the sake of simplicity of notations). Moreover, as $(\varepsilon,\alpha)\to(0,0)$, we have

$$\text{(A.2)}\qquad \begin{cases} \text{(i)}\ \ \bar\mu = \dfrac{w(\bar x,\bar s)-\phi(\bar z)+2\eta(\bar s-\bar t)}{w(\bar y,\bar t)-\phi(\bar z)} \to \mu_\eta(t), \\[2ex] \text{(ii)}\ \ \dfrac{|\bar x-\bar y|^2}{2\varepsilon},\quad \dfrac{|\bar x-\bar z|^2}{2\varepsilon}\to 0,\ \ \alpha|\bar x|^2\to 0, \\[2ex] \text{and} \\[2ex] \text{(iii)}\ \ \bar s>\bar t \text{ and } \bar x,\bar y,\bar z\in\Omega \text{ for } \varepsilon \text{ and } \alpha \text{ small enough,} \end{cases}$$

with the last point being a consequence of the inequality $\mu_\eta(t)<\chi_\eta(t)$.

3. Set

$$\text{(A.3)}\qquad P=\frac{1}{\bar\mu}\frac{(\bar y-\bar x)}{\varepsilon}(w(\bar y,\bar t)-\phi(\bar z)) \ \text{ and }\ Q=\frac{1}{1-\bar\mu}\frac{(\bar z-\bar x)}{\varepsilon}(w(\bar y,\bar t)-\phi(\bar z)).$$

The viscosity inequalities for $w(x,s)$, $w(y,t)$, and ϕ are

$$\text{(A.4)}\qquad \begin{cases} \text{(i)}\ \ -2\eta+H(\bar x,\bar\mu P+(1-\bar\mu)Q+2\alpha\bar x(w(\bar y,\bar t)-\phi(\bar z)))\geq 0, \\[2ex] \text{(ii)}\ \ -\widetilde{\psi}'(\bar t)(w(\bar y,\bar t)-\phi(\bar z))-2\eta\bar\mu^{-1}+H(\bar y,P)\leq 0, \\[2ex] \text{and} \\[2ex] \text{(iii)}\ \ H(\bar z,Q)\leq 0. \end{cases}$$

Using (A.2(ii)) and (A.2(iii)), we may rewrite (A.4) as

$$\text{(A.5)}\qquad \begin{cases} \text{(i)}\ \ -2\eta+H(\bar z,\bar\mu P+(1-\bar\mu)Q)+\tilde n_\varepsilon(\alpha)+\xi(\varepsilon,\alpha)\geq 0, \\[2ex] \text{(ii)}\ \ -\widetilde{\psi}'(\bar t)(w(\bar y,\bar t)-\phi(\bar z))-2\eta\bar\mu^{-1}+H(\bar z,P)-\xi(\varepsilon,\alpha)\leq 0, \\[2ex] \text{and} \\[2ex] \text{(iii)}\ \ H(\bar z,Q)\leq 0, \end{cases}$$

where $\tilde n_\varepsilon(\alpha)\to 0$ when $\alpha\to 0$ if ε is fixed and $\xi(\varepsilon,\alpha)\to 0$ when $(\varepsilon,\alpha)\to(0,0)$.

4. Set

$$\widetilde{P}=\bar\mu(P-Q).$$

If ε and α are chosen sufficiently small and α is small compared to ε, then (A.5(i)) yields

$$H(\bar z,\widetilde{P}+Q)\geq\eta\,,$$

while (A.5(iii)) reads

$$H(\bar{z}, Q) \leq 0.$$

Moreover, again if ε and α are chosen small enough, (A.2(i)) implies that $0 < \bar{\mu} < 1$. Assumption (H4) with $\mu = \bar{\mu}$ then yields

$$\bar{\mu} H(\bar{z}, P) \geq H(\bar{z}, \widetilde{P} + Q) + \psi(\eta)(1 - \bar{\mu}). \tag{A.6}$$

Dividing (A.5(i)) by $\bar{\mu}$ and subtracting (A.5(ii)) we obtain

$$\widetilde{\psi}'(\bar{t})(w(\bar{y}, \bar{t}) - \phi(\bar{z})) + \frac{1}{\bar{\mu}} H(\bar{z}, \widetilde{P} + Q) - H(\bar{z}, P) \geq -\tilde{n}_\varepsilon(\alpha) - \xi(\varepsilon, \alpha)\Big(\frac{1}{\bar{\mu}} + 1\Big), \tag{A.7}$$

and, in view of (A.6),

$$\widetilde{\psi}'(\bar{t})(w(\bar{y}, \bar{t}) - \phi(\bar{z})) + \frac{\psi(\eta)(\bar{\mu} - 1)}{\bar{\mu}} \geq -\tilde{n}_\varepsilon(\alpha) - \xi(\varepsilon, \alpha)\Big(\frac{1}{\bar{\mu}} + 1\Big). \tag{A.8}$$

Dividing by $w(\bar{y}, \bar{t}) - \phi(\bar{z}) \geq 1$, and letting $\alpha \to 0$ and then $\varepsilon \to 0$ we obtain

$$\widetilde{\psi}'(t) + C\psi(\eta) \frac{(\mu_\eta(t) - 1)}{\mu_\eta(t)} \geq 0.$$

Since $0 \leq \mu_\eta(t) \leq 1$, this reduces to

$$\widetilde{\psi}'(t) + C\psi(\eta)(\mu_\eta(t) - 1) \geq 0. \qquad \square$$

We continue with the following proof.

Proof of Theorem 3.2. Since the variational inequality (A.9) below admits a comparison principle, the conclusion follows immediately from the lemma which is stated and proved below. □

LEMMA A.2. *Under the assumptions of Theorem* 2.2, *the function* M_η *defined by* (3.6) *is a viscosity subsolution of the variational inequality*

$$\min[M' + \psi(\eta) M, M - X_\eta] \leq 0 \quad \textit{in } (0, \infty). \tag{A.9}$$

Proof. 1. M_η is clearly positive, as it can be seen by letting $s = t$, uniformly continuous, and bounded, since $w \in BUC(\mathbb{R}^N \times (0, \infty))$.

2. Let $\Phi \in C^1((0, \infty))$ and assume that τ is a local maximum point of $M_\eta - \Phi$ in $[\tau - \delta, \tau + \delta]$ for some $\delta > 0$. Since there is nothing to show if $M_\eta(\tau) \leq X_\eta(\tau)$, we may assume that $M_\eta(\tau) > X_\eta(\tau) \geq 0$.

3. Consider, for $x, y \in \mathbb{R}^N$, $t \in [\tau - \delta, \tau + \delta]$, and $s \geq t$, the function

$$\Psi^{\varepsilon,\alpha}(x, y, t, s) = w(x, t) - w(y, s) - \frac{|x - y|^2}{2\varepsilon^2} - 2\eta(s - t) - \alpha(|x|^2 + |y|^2) - \Phi(t).$$

Classical arguments from the theory of viscosity solutions yield (see [3]) that the function $\Psi^{\varepsilon,\alpha}$ achieves its maximum at some point $(\bar{x}, \bar{y}, \bar{t}, \bar{s})$ and that, when $(\varepsilon, \alpha) \to (0,0)$,

$$\text{(A.10)} \qquad \begin{cases} \text{(i)} \;\; \Psi^{\varepsilon,\alpha}(\bar{x}, \bar{y}, \bar{t}, \bar{s}) \to M_\eta(\tau), \\[2ex] \text{(ii)} \;\; \alpha(|\bar{x}|^2 + |\bar{y}|^2) \to 0, \; \dfrac{|\bar{x} - \bar{y}|^2}{\varepsilon^2} \to 0, \\[2ex] \text{(iii)} \;\; w(\bar{x}, \bar{t}) - w(\bar{y}, \bar{s}) > M_\eta(\bar{t}), \\[2ex] \text{and} \\[2ex] \text{(iv)} \;\; \bar{x}, \bar{y} \in \Omega \text{ and } |\bar{t} - \bar{s}| > 0, \text{ for } (\varepsilon, \alpha) \text{ small, since} \\ \qquad M_\eta(\tau) > X_\eta(\tau) \geq 0. \end{cases}$$

Using (H12) and (H13) we may rewrite the viscosity inequalities

$$\Phi'(\bar{t}) - 2\eta + F(\bar{x}, w(\bar{x}, \bar{t}),\ p + 2\alpha\bar{x}) \leq 0 \quad \text{and} \quad -2\eta + F(\bar{y}, w(\bar{y}, \bar{s}),\ p - 2\alpha\bar{y}) \geq 0,$$

where $p = \frac{(\bar{x} - \bar{y})}{\varepsilon^2}$ as

$$\text{(A.11)} \begin{cases} \text{(i)} \;\; \Phi'(\bar{t}) - 2\eta + F(\bar{x}, w(\bar{x}, \bar{t}), \bar{p}) + \tilde{n}_\varepsilon(2\alpha|\bar{x}|) \leq 0 \\[2ex] \text{and} \\[2ex] \text{(ii)} \;\; -2\eta + F(\bar{x}, w(\bar{y}, \bar{s}), \bar{p}) + m_R(|\bar{x} - \bar{y}|(1 + |p|)) + \tilde{n}_\varepsilon(2\alpha|\bar{y}|) \geq 0. \end{cases}$$

Using (A.10(ii)) we obtain that $F(\bar{x}, w(\bar{y}, \bar{s}), \bar{p}) > \eta$ for α and ε small enough. Since $w(\bar{x}, \bar{t}) \geq w(\bar{y}, s)$, using sufficiently small ε and α in (A.10(iii)) and (H10), yields

$$F(\bar{x}, w(\bar{x}, \bar{t}), \bar{p}) - F(\bar{x}, w(\bar{y}, \bar{s}), \bar{p}) \geq \psi(\eta)(w(\bar{x}, \bar{t}) - w(\bar{y}, \bar{s})) \geq \psi(\eta) M_\eta(\bar{t}).$$

Finally, subtracting (A.11(ii)) for (A.11(i)) we obtain

$$\Phi'(\bar{t}) + \psi(\eta) M_\eta(\bar{t}) + \tilde{n}_\varepsilon(\alpha) + \widetilde{m}(\varepsilon) \leq 0,$$

and we conclude, letting first $\alpha \to 0$ and then $\varepsilon \to 0$. □

REFERENCES

[1] M. Arisawa, *Ergodic problem for the Hamilton-Jacobi-Bellman equation.* I. *Existence of the ergodic attractor*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 415–438.

[2] M. Arisawa, *Ergodic problem for the Hamilton-Jacobi-Bellman equation.* II, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 1–24.

[3] G. Barles, *Asymptotic behavior of viscosity solutions of first-order Hamilton-Jacobi equations*, Ricerche Mat., 34 (1985), pp. 227–260.

[4] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[5] A. Fathi, *Sur la convergence du semi-groupe de Lax–Oleinik*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 267–270.

[6] P.-L. Lions, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, Boston, MA, 1982.

[7] P.-L. Lions, G. Papanicolaou, and S. R. S. Varadhan, *Homogenization of Hamilton–Jacobi Equations*, preprint.

[8] G. Namah and J. M. Roquejoffre, *Remarks on the long time behavior of the solutions of Hamilton–Jacobi Equations*, Comm. Partial Differential Equations, 24 (1999), pp. 883–893.

[9] J. M. Roquejoffre, *Comportement asymptotique des solutions d'équations de Hamilton–Jacobi monodimensionnelles*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 185–189.

[10] J. M. Roquejoffre, *Convergence to Steady States or Periodic Solutions in a Class of Hamilton–Jacobi Equations*, preprint.

SIAM J. MATH. ANAL.
Vol. 31, No. 4, pp. 940–940

THE SHAPE OF THE TALLEST COLUMN: CORRECTED*

STEVEN J. COX† AND C. MAEVE MCCARTHY‡

Abstract. Our summary of the work of Keller and Niordson (*J. Math. Mech.* 16, pp. 433–446, 1966) was inaccurate. We offer a correction.

Key words. isolated eigenvalue, continuous spectrum

AMS subject classifications. 35L15, 49J99, 73H05

PII. S0036141099354958

Our criticism [1] of the work of Keller and Niordson [2] was not accurate. Following Proposition 2.1 on page 549 of [1] we stated that "Keller and Niordson's calculation of $c = \lambda_1/24$ suggests that their design gives rise to an isolated eigenvalue, $\lambda_1 = 24c$, just below the continuous spectrum." We followed this statement with the misrepresentation "Their result, however, was predicated on the false assumption that (2.4) possessed a purely discrete spectrum." We wish to replace this statement with "Their result, however, assumed the existence, that cannot be taken for granted, of such an isolated eigenvalue for a in a neighborhood of the optimal design." We regret the misrepresentation.

REFERENCES

[1] S.J. Cox and C.M. McCarthy, *The shape of the tallest column*, SIAM J. Math. Anal., 29 (1998), pp. 547–554.

[2] J.B. Keller and F.I. Niordson, *The tallest column*, J. Math. Mech., 16 (1966), pp. 433–446.

*Received by the editors April 26, 1999; accepted for publication June 11, 1999; published electronically April 4, 2000.

http://www.siam.org/journals/sima/31-4/35495.html

†Department of Computational and Applied Mathematics, Rice University, 6100 Main St., Houston, TX 77005 (cox@rice.edu).

‡Department of Mathematics and Statistics, Murray State University, Murray, KY 42071 (maeve.mccarthy@murraystate.edu).

SIAM J. Math. Anal.
Vol. 31, No. 5, pp. 941–991

NONCLASSICAL SHOCKS AND KINETIC RELATIONS: STRICTLY HYPERBOLIC SYSTEMS*

BRIAN T. HAYES[†] AND PHILIPPE G. LEFLOCH[‡]

Abstract. We consider strictly hyperbolic systems of conservation laws whose characteristic fields are not genuinely nonlinear, and we introduce a framework for the nonclassical shocks generated by diffusive or diffusive-dispersive approximations. A nonclassical shock does not fulfill the Liu entropy criterion and turns out to be undercompressive.

We study the Riemann problem in the class of solutions satisfying a single entropy inequality, the only such constraint available for general diffusive-dispersive approximations. Each nongenuinely nonlinear characteristic field admits a *two-dimensional wave set*, instead of the classical one-dimensional wave curve. In specific applications, these wave sets are narrow and resemble the classical curves. We find that even in strictly hyperbolic systems, nonclassical shocks with arbitrarily small amplitudes occur. The Riemann problem can be solved uniquely using nonclassical shocks, provided an additional constraint is imposed: we stipulate that the entropy dissipation across any nonclassical shock be a given constitutive function. We call this admissibility criterion a *kinetic relation*, by analogy with similar laws introduced in material science for propagating phase boundaries. In particular, the kinetic relation may be expressed as a function of the propagation speed. It is derived from traveling waves and, typically, depends on the ratio of the diffusion and dispersion parameters.

Key words. conservation laws, hyperbolic entropy, shock wave, kinetic relation, nonclassical shock

AMS subject classifications. 35L65, 76L05

PII. S0036141097319826

1. Introduction.

In this paper, we consider discontinuous solutions to hyperbolic systems of conservation laws that do not fulfill the classical entropy criteria, carrying over to systems the discussion we initiated in [22] for scalar equations with nonconvex fluxes. We develop a framework for the existence and uniqueness of the *nonclassical* shock waves that arise as limits of diffusive-dispersive approximations. It is natural to constrain the solutions to the hyperbolic system with an entropy inequality for a *single*, strictly convex entropy pair. This condition is weaker than the Liu [41] entropy criterion.

A *nonclassical shock* is defined as one that does not satisfy the Liu criterion. It turns out that such a shock is *undercompressive*: the number of characteristics impinging on the discontinuity is smaller than that imposed by the (classical) Lax shock inequalities. Such waves are underdetermined (in the sense of linear analysis) and sensitive to the form of the diffusive-dispersive mechanism.

The focus of this work is on strictly hyperbolic systems where one (or more) characteristic field lacks genuine nonlinearity, such as those describing the dynamics of elastic materials or magnetic fluids. A key observation is that undercompressive shocks may arise for such systems through balanced diffusive and dispersive mechanisms: this

*Received by the editors April 16, 1997; accepted for publication (in revised form) February 24, 1998; published electronically April 20, 2000.

http://www.siam.org/journals/sima/31-5/31982.html

[†]Department for Mathematics, Duke University, Durham, NC 27706 (bhayes@math.duke.edu).

[‡]Centre de Mathématiques Appliquées and Centre National de la Recherche Scientifique, UA 756, Ecole Polytechnique, 91128 Palaiseau Cedex, France (lefloch@cmapx.polytechnique.fr). This author was partially supported by the Centre National de la Recherche Scientifique, the National Science Foundation under grants DMS 94-01003 and DMS 95-02766 and a Faculty Early Career Development (CAREER) award, and the Air Force Office under grant F49620-94–1-0215.

is the case even for shocks having *arbitrary small* amplitude. We concentrate here on the Riemann problem which is fundamental in the theory of conservation laws. A typical Riemann solution combines classical (shock and rarefaction) waves and nonclassical shocks. The numerical analysis of nonclassical shocks is investigated in a companion paper [23].

We build here upon extensive activity on undercompressive waves for nonstrictly hyperbolic systems and systems with change of type. In the examples studied in the literature, the undercompressive waves have finite strength; they were found to be necessary in order to solve the Riemann problem and, therefore, reflect a property of the flux-function of the system. We refer the reader to Azevedo et al. [3], Freistühler [18], Isaacson, Marchesin, and Plohr [28], Isaacson et al. [27], Keyfitz [31], Liu and Zumbrun [46, 47], Schecter and Shearer [52], Slemrod [59], and the references therein.

The basic concepts and the analysis of the traveling waves associated with such nonstandard discontinuities and a resolution of the Riemann problem for some mathematical models can be also found in [28, 32, 44, 45, 58]. The large-time asymptotic stability of under- or overcompressive shocks (the number of impinging characteristics in the latter is larger) is proven in [19, 43, 46, 47]. Liu and Zumbrun observe [47] that, for undercompressive shocks, the asymptotic state for large times cannot be determined solely from the mass of the initial perturbation, but must also take into account the diffusive effects of a parabolic augmented system of equations.

Several examples from continuum mechanics are known to exhibit undercompressive shocks. The system of magnetohydrodynamics lacks both genuine nonlinearity and strict hyperbolicity (Brio and Wu [5]). It has been observed numerically, as well as analytically, that nonstandard shock waves not fulfilling the classical entropy criteria arise with certain approximations.

MHD shocks may be either undercompressive or overcompressive. Those shocks are called nonstandard or intermediate in the MHD literature and are critical to the understanding of important phenomena such as the effect of the solar wind (Wu [63]). For various results on the Riemann problem for a rotationally invariant model in MHD, we refer to [4, 6, 8, 17, 32, 65]. See also [21] for another model. There is also an extensive literature on phase boundaries in materials admitting phase transformations of the austenite-martensite type. When the stress-strain relation for a material is decreasing on an interval, the system of elastodynamics is of the hyperbolic-elliptic type. Propagating phase boundaries are still another example of undercompressive waves. They are fundamental to understanding phase transformation processes. See [15, 26, 56, 57, 59] as well as [1, 2, 38, 61, 62]. See also [48] for a general review on the nonlinear waves arising in fluids and materials, with or without phase transitions.

A pioneering study of the effect of vanishing diffusion and dispersion terms in scalar conservation laws can be found in Schonbek [53] using the compensated compactness method. She proved a convergence theorem toward weak solutions. LeFloch and Natalini [39] used the concept of measure-valued solution and established convergence results assuming that the diffusion dominates the dispersion.

The works by Wu [64] and Jacobs, McKinney, and Shearer [30] established the first existence result of undercompressive shocks for the modified KdV–Burgers equation and motivated us in [22].

The present series of papers [22, 23] is intended as a contribution toward unifying ideas behind some of the above works. We pursue a better understanding of simple models giving rise to undercompressive shocks. Deriving entropy criteria for their selection is one of the main challenges in the field. The classical criteria developed by Dafermos [10, 11, 12], Lax [34, 35], Liu [41, 42], Oleinik [50], etc., cannot be applied directly. In contrast to previous works, we focus here primarily on strictly hyperbolic systems having nongenuinely nonlinear characteristic fields.

Given a strictly convex entropy pair, we first endeavor to describe the set of all solutions to the Riemann problem that satisfy a single entropy inequality. Allowing nonclassical shocks leads to a lack of uniqueness for the Riemann problem and a *multi-parameter family* of solutions can be constructed. Our construction is an extension to Liu's theorem on the resolution of the Riemann problem which was based on what is now called the Liu criterion. This analysis provides a complete description of all the Riemann solutions generated by any diffusive-dispersive approximation compatible with a given entropy pair (section 2). We observe that characterizing limits of approximate sequences of solutions to hyperbolic systems via pointwise relations on the propagating discontinuities in the limiting solution may not be possible in the most general situation (see, for instance, Glimm [20] and LeFloch and Tzavaras [40]). In this regard, our analysis is pertinent toward describing the *set* of all possible such limits. In our presentation, pointwise constraints are added afterward.

Next we investigate a way of selecting a unique nonclassical solution. We propose to make the selection based on the entropy dissipation, which is a fundamental quantity from both mathematical and physical standpoints. We stipulate that the entropy dissipation of a nonclassical shock be a given function, the "kinetic function." It may be assumed, for instance, that the kinetic function depends only on the *speed of the nonclassical shock*. We call such an admissibility criterion a *kinetic relation* by analogy with similar laws introduced in material science.

Therefore this generalizes to strictly hyperbolic systems the notion of kinetic relation known for the hyperbolic-elliptic system of phase transitions (Abeyaratne and Knowles [1, 2] and Truskinovsky [61, 62]; see also LeFloch [38]) and for nonconvex scalar conservation laws (Hayes and LeFloch [22] and Kulikovsky [33]). The paper by Truskinovsky [62] includes a review of these issues in material science.

In section 2, we construct a unique solution to the Riemann problem in the class of nonclassical solutions when the kinetic relation is enforced. For some Riemann data choosing between the classical solution and the nonclassical one may be still necessary (see section 2). When a specific augmented system including diffusive/dispersive effects is provided, the entropy dissipation and therefore the kinetic function can be determined. Small-scale effects neglected in the mathematical modeling at the hyperbolic level are essential to understanding the behavior of nonclassical shocks. The kinetic function can be obtained from the equation of the traveling wave solutions associated with the diffusive-dispersive model.

Classical and nonclassical shock are very different in nature. The classical shocks are associated with the continuum spectrum of the traveling wave equation and the nonclassical shocks with its discrete spectrum. Typically, given a (left) state, and restricting attention to a given wave family, there exists a one-parameter family of right states that can be attained with a classical shock, but a single right state can be attained by a nonclassical shock.

In several systems arising in the applications in continuum mechanics, the entropy dissipation is related to the total energy and may be viewed as a force driving the propagation of the nonclassical propagating discontinuities. We also consider here the Riemann problems with large amplitude for two specific examples of interest: a system from nonlinear elastodynamics based on a nonconvex strain-stress law, which is a strictly hyperbolic system with two nongenuinely nonlinear fields (sections 3 and 4), and a model from magnetohydrodynamics, which has an umbilic point and one linearly degenerate characteristic field (section 5). In these examples we demonstrate numerically that certain diffusive-dispersive approximations generate nonclassical shocks.

The kinetic relation may be used in the design of a numerical scheme consistent with the underlying regularization, avoiding the (costly) resolution of small-scale effects. Hou, LeFloch, and Rosakis [25] proposed recently, for computing propagating phase boundaries in a two-dimensional plate, a consistent method based on the level set formulation. For difference schemes generating nonclassical shocks, one can consult [23, 24].

2. A framework for nonclassical shocks in systems.

2.1. Preliminaries. Here we shall motivate the definition of nonclassical solution. Consider a system of hyperbolic conservation laws:

$$\partial_t u + \partial_x f(u) = 0, \qquad u(x,t) \in \mathcal{U}, \tag{2.1}$$

where $\mathcal{U}$ is a convex and open subset of $\mathbb{R}^N$ and the flux-function $f : \mathcal{U} \to \mathbb{R}^N$ is a smooth mapping. We assume that the system is endowed with a strictly convex entropy pair (U, F); that is, $\nabla F^T = \nabla U^T Df$ and $\nabla^2 U(u) \geq C\, Id$ with $C > 0$. This, in particular, implies that the system is hyperbolic, although not necessarily strictly hyperbolic.

Suppose that the "good" solutions to (2.1) according to some underlying physical interpretation are to be obtained as limits of a diffusive-dispersive approximation scheme of the form

$$\partial_t u_\epsilon + \partial_x f(u_\epsilon) = \epsilon\, \partial_x\big(B_1(u_\epsilon)\partial_x u_\epsilon\big) + \epsilon^2\, \partial_x\big(B_2(u_\epsilon)\partial_{xx} u_\epsilon\big) \tag{2.2}$$

as $\epsilon \to 0$ ($\epsilon > 0$). When B_1 and B_2 are $N \times N$ matrix-valued functions, the regularization (2.2) (together with the conditions (2.3) below) describes one large class of systems, which includes the examples in the applications we will be interested in. (The important issue of the existence of a solution u_ϵ satisfying (2.2) is out of the scope of the present paper.)

We shall say that the pair (U, F) is *compatible* with the approximation scheme (2.2) if the following conditions hold:

• The first term in the right-hand side of (2.2) is *dissipative* for the entropy U, in the sense that

$$\nabla^2 U(v) B_1(v) \text{ is a positive matrix for all } v \in \mathcal{U}. \tag{2.3i}$$

• The second term in (2.2) is *conservative* for U, in the sense that there exist $N \times N$ matrix-valued functions B_3 and B_4 such that

$$\partial_x v^T \nabla^2 U(v)^T B_2(v) \partial_{xx} v \;=\; \partial_t\big(\partial_x v^T B_3(v) \partial_x v\big) \;+\; \partial_x\big(\partial_x v^T B_4(v) \partial_x v\big) \tag{2.3ii}$$

for any solution $v : \mathbb{R} \times \mathbb{R}_+ \to \mathcal{U}$ to (2.2), and

$$B_3(v) \text{ is a nonnegative matrix for all } v \in \mathcal{U}. \tag{2.3iii}$$

Note in passing that trivial linear entropies always satisfy (2.3) but are of no use for our purpose of selecting solutions to (2.1). When (2.3) holds and $\partial_x u_\epsilon$ vanishes at infinity, one can (formally) derive from (2.2) an entropy inequality. Indeed we obtain

$$\begin{aligned}\partial_t\big(U(u_\epsilon) + {}& \epsilon^{\,2}\,\partial_x u_\epsilon^T B_3(u_\epsilon)\partial_x u_\epsilon\big) + \partial_x F(u_\epsilon)\\ ={}& \epsilon\,\partial_x\big(\nabla U(u_\epsilon)^T B_1(u_\epsilon)\partial_x u_\epsilon\big) - \epsilon\,\partial_x u_\epsilon \nabla^2 U(u_\epsilon) B_1(u_\epsilon)\partial_x u_\epsilon\\ &+ \epsilon^2\,\partial_x\big(\nabla U(u_\epsilon)^T B_2(u_\epsilon)\partial_{xx} u_\epsilon\big) - \epsilon^2\,\partial_x\big(\partial_x u_\epsilon B_4(u_\epsilon)\partial_x u_\epsilon\big),\end{aligned}$$

which yields the balance law

$$\begin{aligned}\int_{\mathbb{R}} U(u_\epsilon(t))\,dx \,+\, \epsilon^2 \int_{\mathbb{R}} &\partial_x u_\epsilon(t)^T B_3(u_\epsilon(t))\partial_x u_\epsilon(t)\,dx\\ &+ \epsilon \int_0^t \int_{\mathbb{R}} \partial_x u_\epsilon^T \nabla^2 U(u_\epsilon) B_1(u_\epsilon)\partial_x u_\epsilon\,dxds\\ &= \int_{\mathbb{R}} U(u_\epsilon(0))\,dx \,+\, \epsilon^2 \int_{\mathbb{R}} \partial_x u_\epsilon(0)^T B_3(u_\epsilon(0))\partial_x u_\epsilon(0)\,dx\end{aligned} \tag{2.4}$$

for all $t \geq 0$ and an entropy inequality for $u = \lim_{\epsilon\to 0} u_\epsilon$ of

$$\partial_t U(u) + \partial_x F(u) \,\leq\, 0. \tag{2.5}$$

We observe that

• an arbitrary entropy for (2.1) need not be compatible with the given regularization (2.2), and the inequality (2.5) *need not hold* for an arbitrary entropy;

• the estimate (2.4) provides an a priori control on u_ϵ and its derivatives, which may be used to apply the compensated compactness method, at least if $N \leq 2$. When the latter applies the sequence u_ϵ is shown to converge to a weak solution to (2.1), (2.5). See [53, 22] and sections 4 and 5 of this paper.

As an illustration, consider the case of a scalar equation ($N = 1$) and

$$\partial_t u_\epsilon + \partial_x f(u_\epsilon) = \epsilon\,\partial_{xx} u_\epsilon + \alpha\,\epsilon^2\,\partial_{xxx} u_\epsilon, \tag{2.6}$$

where α is a real parameter. It is easily checked that the conditions (2.3) hold for $U(u) = u^2$ with $B_1 = 1$, $B_2 = \alpha$, $B_3 = \alpha/2$, and $B_4 = 0$. The estimate (2.4) reduces to

$$\int_{\mathbb{R}} u_\epsilon(t)^2\,dx \,+\, 2\,\epsilon \int_0^T \int_{\mathbb{R}} |\partial_x u_\epsilon|^2\,dxds \,=\, \int_{\mathbb{R}} u_\epsilon(0)^2\,dx, \tag{2.7}$$

and we get the inequality

$$\partial_t U(u) + \partial_x F(u) \,\leq\, 0, \qquad F'(u) := u\,f'(u).$$

Observe that, for nonquadratic entropies, (2.3) is generally violated and the inequality (2.5) does not hold, as was pointed out in Hayes and LeFloch [22].

The scaling in (2.6) is important. The diffusion dominant regularization

$$\partial_t u_\epsilon + \partial_x f(u_\epsilon) = \epsilon\,\partial_{xx} u_\epsilon + \delta\,\partial_{xxx} u_\epsilon \tag{2.8}$$

with $\delta = o(\epsilon^2)$ would bring us back to the classical theory of conservation laws, while the dispersion dominant case (2.8) with $\epsilon^2 = o(\delta)$ is the subject of the Lax–Levermore theory [36, 37]. Limiting solutions in the latter case are not weak solutions to (2.1).

This motivates us to constrain the solutions to (2.1) with the single entropy inequality (2.5). Not surprisingly, when one characteristic field (or more) of the system

(2.1) is not genuinely nonlinear, the entropy inequality will be shown to be too lax to guarantee uniqueness even for the Riemann problem. The forthcoming analysis is built upon this elementary observation.

Our analysis in [22] of the nonclassical shocks for scalar conservation laws relied on the violation of the Oleinik criterion. For systems we shall say that a shock is classical if it satisfies the Liu criterion. Definition 2.1 restates this concept.

Definition 2.1. *A propagating discontinuity is called a* nonclassical shock *when it satisfies the entropy inequality* (2.5) *but does not fulfill the Liu entropy criterion (see* (2.18) *below).* □

2.2. Nonclassical Riemann solutions. We now study the Riemann problem for nongenuinely nonlinear systems.

- Liu has constructed a unique entropy solution to the Riemann problem for such systems [41, 42]. In his construction, every shock satisfies what is now called the Liu criterion. This is described in Lemma 2.3.
- When a single entropy inequality is used, the class of admissible solutions is larger (Lemma 2.5) and undercompressive shocks are found near a curve where genuine nonlinearity breaks down (see Lemma 2.4).
- We construct a multiparameter family of solutions to the Riemann problem in Theorem 2.6. In our construction, there are two analogous cases corresponding to a minimum or a maximum of the wave speed at the point where genuine nonlinearity is lost.

This extends Liu's construction to encompass all possible limits of diffusive-dispersive approximations compatible with a given entropy pair (U, F). A further admissibility criterion will be necessary to ensure uniqueness of the entropy solution. This will be developed in subsection 2.3.

Remark 2.2. Liu's criterion is consistent with the regularization (2.2) with $B_1(u) = I$ and $B_2(u) = 0$. The latter regularization happens to be compatible with *any* convex entropy to (2.1) since, then, (2.3i) is equivalent to the convexity assumption on U and (2.3ii) and (2.3iii) are trivially satisfied. Henceforth the inequalities (2.5) in this particular case hold for *all* convex entropy pairs. However, the Liu criterion need not be satisfied by limits of more general diffusive approximations or by diffusive-dispersive ones.

We now assume that $\mathcal{U} := B(u_*, R)$ is a ball with center u_* and radius $R > 0$, and, for each u and u' in $\mathcal{U}$, the matrix $A(u, u') := \int_0^1 Df(m\,u + (1-m)\,u')\,dm$ admits N real and distinct eigenvalues $\bar\lambda_1(u, u') < \bar\lambda_2(u, u') < \cdots < \bar\lambda_N(u, u')$ and corresponding basis of right eigenvectors $\bar r_j(u, u')$ and left eigenvectors $\bar l_j(u, u')$. Throughout this section we normalize the basis so that $\bar l_j(u, u') \cdot \bar r_j(u, u') = \delta_{ij}$.

It is assumed that the wave speeds $\lambda_j(u, u')$ are strictly separated in the sense that there exist *disjoint* intervals $\big[\lambda_j^{\min}, \lambda_j^{\max}\big]$, $j = 1, 2, \ldots, N$, such that

$$\lambda_j^{\min} < \bar\lambda_j(u, u') < \lambda_j^{\max} \tag{2.9}$$

for all $u, u' \in \mathcal{U}$. We also set $\lambda_j(u) := \bar\lambda_j(u, u)$, $r_j(u) := \bar r_j(u, u)$, and $l_j(u) := \bar l_j(u, u)$. When (2.1) is strictly hyperbolic, the condition (2.9) is satisfied if $\mathcal{U}$ is a sufficiently small neighborhood of u_*.

We are interested in systems admitting $N - P$ genuinely nonlinear characteristic fields and $P \le N$ nongenuinely nonlinear characteristic fields. In the latter case the scalar-valued function $u \to \nabla\lambda_j(u) \cdot r_j(u)$ does not keep a constant sign. We assume that there is a subset with P elements, $\mathbf{P} \subset \big\{1, 2, \ldots, N\big\}$ such that, for $j \notin \mathbf{P}$,

$\nabla\lambda_j(u)\cdot r_j(u) > 0$ for all u (after suitable normalization of the eigenvectors), and for $j \in \mathbf{P}$, the set

$$\mathcal{M}_j = \big\{u \in \mathcal{U} \,|\, \nabla\lambda_j(u)\cdot r_j(u) = 0\big\}$$

is a smooth affine manifold with dimension $N-1$ containing the point u_*. For simplicity in the presentation we do not include linearly degenerate fields.

We denote by $\mu_j(u)$ a scalar-valued function satisfying $\nabla\mu_j \cdot r_j \equiv 1$. When the j-field is genuinely nonlinear, one takes $\mu_j(u) = \lambda_j(u)$. The function μ_j will be used to parameterize the wave curves. We assume that μ_j can be chosen such that

$$\mu_j(u) = 0 \qquad \text{iff} \qquad \nabla\lambda_j(u)\cdot r_j(u) = 0,$$

and either

$$\text{(2.10a)} \qquad \text{Case A: } \mu_j(u) \quad \text{and} \quad \nabla\lambda_j(u)\cdot r_j(u) \quad \text{have the same sign,}$$

or

$$\text{(2.10b)} \qquad \text{Case B: } \mu_j(u) \quad \text{and} \quad \nabla\lambda_j(u)\cdot r_j(u) \quad \text{have the opposite sign.}$$

In particular $\nabla\lambda_j \cdot r_j$ changes sign across $\mathcal{M}_j$. In Case A, $\mu_j(u) = 0$ is associated with a minimum of the wave speed, while in Case B it is associated with a maximum. In the scalar case, (2.10a) means that there is a state u_* such that the function f is strictly concave for $u < u_*$ and strictly convex for $u > u_*$. Typical examples are $f(u) = u^3$ in the case (2.10a) and $f(u) = -u^3$ in the case (2.10b); in both cases one can choose $\mu(u) = u$. As we will see, the cases (2.10a) and (2.10b) lead to wave curves with different properties.

The Riemann problem, (2.1) with initial data

$$\text{(2.11)} \qquad u(x,0) = \begin{cases} u_l & \text{for } x < 0, \\ u_r & \text{for } x > 0, \end{cases}$$

and u_r and u_l fixed in $\mathcal{U}$, plays an important role in the theory of hyperbolic conservation laws. Since the problem is invariant under the transformation $(x,t) \to (\beta\, x, \beta\, t)$ (with $\beta > 0$), it is natural to search for self-similar solutions depending only on x/t. We now define the one-parameter families of shock and rarefaction waves to be used as building blocks in the resolution of the Riemann problem.

Given a state $u_0 \in \mathcal{U}$ and $j = 1, 2, \dots, N$, let $\mathcal{O}_j(u_0) = \big\{v_j(\epsilon_j; u_0) \in \mathcal{U}\big\}$ be the integral curve of the vector field r_j issued from u_0, so that

$$\text{(2.12)} \qquad \frac{dv_j}{d\epsilon_j}(\epsilon_j; u_0) = r_j\big(v_j(\epsilon_j; u_0)\big), \qquad v_j(\epsilon_{j,0}; u_0) = u_0.$$

Note that $r_j(u_0)$ is the tangent vector of the curve $\mathcal{O}_j(u_0)$ at the point u_0. Using the normalization of the function μ_j, one checks that

$$\mu_j\big(v_j(\epsilon_j; u_0)\big) = \epsilon_j;$$

therefore there should be no confusion in using the notation $mu_j = \epsilon_j$. In other words, μ_j is viewed as both a function of u and as a parameter along the wave curves.

We also consider the Hugoniot locus

$$\mathcal{H}_j(u_0) := \big\{ w \,|\, -s\,\big(w - u_0\big) + f(w) - f(u_0) = 0 \big\}. \tag{2.13}$$

The Rankine–Hugoniot relation is equivalent to saying that there exists an index j and a scalar-valued coefficient $\alpha(u_0, w)$ such that

$$w - u_0 = \alpha(u_0, w)\, \bar{r}_j(u_0, w), \qquad s = \bar{\lambda}_j(u_0, w). \tag{2.14}$$

By the implicit function theorem, the Hugoniot set decomposes (locally near u_0, at least) into N Hugoniot curves $\mathcal{H}_j(u_0) = \big\{ w_j(\mu_j; u_0) \in \mathcal{U} \big\}$, passing through u_0 and having the tangent vector $r_j(u_0)$ at u_0. Since $\nabla \mu_j \cdot r_j > 0$, the coefficient $\alpha(u_0, w_j)$ in (2.14) has the same sign as that of $\mu_j(w_j) - \mu_j(u_0)$. Along the j-curve, the shock speed satisfies

$$\bar{\lambda}_j(u_0, w_j) = \lambda_j(u_0) + \frac{\mu_j}{2}\, \nabla \lambda_j(u_0) \cdot r_j(u_0) + O(\mu_j^2). \tag{2.15}$$

Taking a suitable subset $B(u_*, R')$ of $\mathcal{U} = B(u_*, R)$ if necessary, one can assume that the curves $\mathcal{O}_j(u_0)$ and $\mathcal{H}_j(u_0)$ extend up to the boundary of $\mathcal{U}$. Furthermore we assume that, for $j \in \mathbf{P}$, these curves are *transverse to the manifold* $\mathcal{M}_j$: each Hugoniot curve and each integral curve intersect the manifold at exactly one point. Observe that when R is small enough, it is sufficient to assume that the vector field r_j is transverse to the manifold $\mathcal{M}_j$. Our construction here applies, however, to the case that R is not necessarily small. The tranversality assumption implies that, for $j \in \mathbf{P}$, the wave speed $\mu_j \to \lambda_j\big(v_j(\mu_j; u_0)\big)$ has exactly one critical point along each integral curve. It will be checked in Lemma 2.3 below that, for $j \in \mathbf{P}$, the shock speed $\mu_j \to \bar{\lambda}_j(u_0, w_j(\mu_j; u_0))$ also admits (at most) one critical point along the Hugoniot curve.

Finally we introduce another assumption about the Hugoniot curve, for all $w_j(\mu_j; u_0)$ with $\mu_j \neq \mu_j(u_0)$,

$$l_j(w_j) \cdot \frac{dw_j}{d\mu_j} \;>\; 0, \tag{2.16i}$$

$$\big(\mu_j - \mu_j(u_0)\big)\, l_j(w_j) \cdot (w_j - u_0) \;>\; 0. \tag{2.16ii}$$

Both conditions in (2.16) trivially hold for weak shocks, since $l_j(u_0) \cdot r_j(u_0) = 1$.

Discontinuous solutions being not unique in general, it is customary to select the "admissible" weak solutions via an entropy criterion acting on discontinuities. From physical, mathematical, and numerical standpoints, it is desirable that an admissible solution to the Riemann problem exist, be unique, and depend continuously upon its initial states in a certain topology. In the classical approach, a wave curve $\mathcal{W}_j(u)$ is indeed defined by piecing together (admissible) parts of the above curves. The Lax shock inequalities [34, 35] are fundamental for stability and are used for weak shocks in the neighborhood of a point of genuine nonlinearity. A j-shock connecting u_0 to u_1 with speed $\lambda_j(u_0, u_1)$ is admissible in the sense of Lax iff

$$\lambda_j(u_0) \geq \bar{\lambda}_j(u_0, u_1) \geq \lambda_j(u_1). \tag{2.17}$$

Note that the inequalities $\lambda_{j-1}(u_0) < \bar{\lambda}_j(u_0, u_1) < \lambda_{j+1}(u_1)$ are obtained as a direct consequence of (2.9). When the characteristic fields are genuinely nonlinear, applying the Lax criterion leads to uniquely defined wave curves and to a unique solution for the Riemann problem. Each wave curve contains two distinct parts, half of the Hugoniot curve and half of the integral curve.

When one or more characteristic fields are not genuinely nonlinear, Liu proposed that, along the Hugoniot curve $\mathcal{H}_j(u_0)$, the following criterion holds:

$$\bar{\lambda}_j(u_0, w_j(\mu_j; u_0)) \geq \bar{\lambda}_j(u_0, u_1) \tag{2.18}$$

for all μ_j between $\mu_j(u_0)$ and $\mu_j(u_1)$; in other words, the shock speed for μ_j in the above range achieves its *minimum* at the point u_1. Liu [42] constructed a unique wave curve based on the condition (2.18). The wave curves may be composed of more than two pieces, and the Riemann solution contains composite waves mixing shocks and rarefactions.

It is known that (2.8), (2.17), and (2.18) are equivalent for shocks of weak amplitude and genuinely nonlinear fields. This is not true for systems having nongenuinely nonlinear fields. In the present paper we attempt to construct a wave curve based on (2.5) of the wave curves of Liu. However, instead of one-parameter wave curves we arrive here to two-parameter sets, which we call “wave sets.” In this construction it is important to distinguish several types of discontinuities.

An arbitrary j-shock connecting u_0 to u_1 can be either *a Lax shock*, in which case (2.17) holds, *an undercompressive shock* satisfying either

$$\bar{\lambda}_j(u_0, u_1) \leq \min\bigl(\lambda_j(u_0), \lambda_j(u_1)\bigr) \text{ or} \tag{2.19}$$

$$\bar{\lambda}_j(u_0, u_1) \geq \max\bigl(\lambda_j(u_0), \lambda_j(u_1)\bigr), \tag{2.20}$$

or *a rarefaction shock*:

$$\lambda_j(u_0) < \bar{\lambda}_j(u_0, u_1) < \lambda_j(u_1). \tag{2.21}$$

The properties of the wave speeds and shock speeds are described in Lemmas 2.3 and 2.4. (See Figure 2.1 for a graphical representation.) The entropy dissipation is dealt with in Lemma 2.5.

LEMMA 2.3. *Let u_0 be given with $\mu_j(u_0) > 0$ and consider the Hugoniot curve $\mathcal{H}_j(u_0)$ for $= j = 1, 2, \dots, N$. Suppose that* (2.10*a*) *(resp.,* (2.10*b*)*) holds. Then the wave speed $\mu_j \to g(\mu_j; u_0) := \lambda_j(w_j(\mu_j; u_0))$ is decreasing (resp., increasing) for $\mu_j < 0$ and increasing (resp., decreasing) for $\mu_j > 0$ and achieves its minimum (resp., maximum) at $\mu_j = 0$.*

There exists $\mu_j^\star(u_0) \leq 0$ such that the shock speed $\mu_j \to h(\mu_j; u_0) := \bar{\lambda}_j(u_0, w_j(\mu_j; u_0))$ is decreasing (resp., increasing) for $\mu_j < \mu_j^\star(u_0)$ and increasing (resp., decreasing) for $\mu_j > \mu_j^\star(u_0)$ and achieves its minimum (resp., maximum) at $\mu_j^\star(u_0)$.

The wave speed and the shock speed coincide at the critical value of the shock speed:

$$g(\mu_j^\star(u_0); u_0) = h(\mu_j^\star(u_0); u_0). \tag{2.22}$$

Moreover we have in case (2.10*a*)

$$\begin{aligned} &h(\mu_j; u_0) - g(\mu_j; u_0) > 0 \qquad \textit{for } \ \mu_j \in \bigl(\mu_j^\star(u_0), \mu_j(u_0)\bigr), \\ &h(\mu_j; u_0) - g(\mu_j; u_0) < 0 \qquad \textit{for } \ \mu_j < \mu_j^\star(u_0) \ \ \textit{or} \ \ \mu_j > \mu_j(u_0), \end{aligned} \tag{2.23a}$$

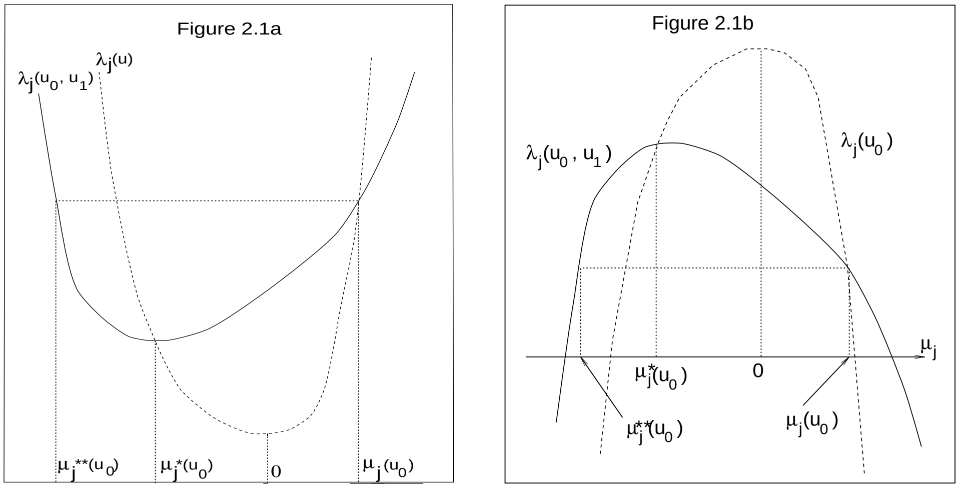


FIG. 2.1. *Wave speed and shock speed for* (a) *the case* (2.10a), (b) *the case* (2.10b).

and in the case (2.10b) *we have*

$$\begin{aligned} &h(\mu_j; u_0) - g(\mu_j; u_0) < 0 \qquad \textit{for } \mu_j \in \big(\mu_j^\star(u_0), \mu_j(u_0)\big), \\ &h(\mu_j; u_0) - g(\mu_j; u_0) > 0 \qquad \textit{for } \mu_j < \mu_j^\star(u_0) \ \textit{ or } \ \mu_j > \mu_j(u_0). \end{aligned} \tag{2.23b}$$

When $\mu_j(u_0) = 0$*, the same properties hold with* $\mu_j^\star(u_0) = 0$*.*

Lemma 2.3 includes, as a special case, the situation that the point $w_j(\mu_j^\star(u_0); u_0)$ belongs to the boundary of $\mathcal{U}$, in which case $\big\{\mu_j < \mu_j^\star(u_0)\big\}$ is empty. We denote by $\mu_j^{\star\star}(u_0)$, with $\mu_j^{\star\star}(u_0) < \mu_j^\star(u_0)$, the point of the Hugoniot curve such that

$$h(\mu_j^{\star\star}(u_0); u_0) = h(\mu_j(u_0); u_0) \tag{2.24}$$

when such a point exists. In the following, we tacitly assume that both points, $\mu_j^\star(u_0)$ and $\mu_j^{\star\star}(u_0)$, exist and belong to the interior of $\mathcal{U}$, the discussion below being much simpler in other cases. Lemma 2.3 is due to Liu [42] and, for completeness, a proof is given in the appendix.

LEMMA 2.4. *Let* u_0 *be given with* $\mu_j(u_0) \geq 0$ *and consider the Hugoniot curve* $\mathcal{H}_j(u_0)$*.*

(1) *Suppose that* (2.10a) *holds. A shock connecting* u_0 *to* $u_1 = w_j(\mu_j(u_1); u_0)$ *is*

$$\begin{aligned} &\textit{a rarefaction shock if } \mu_j(u_1) > \mu_j(u_0) \ \textit{ or } \ \mu_j(u_1) < \mu_j^{\star\star}(u_0); \\ &\textit{a Lax shock if } \mu_j(u_1) \in \big[\mu_j^\star(u_0), \mu_j(u_0)\big]; \\ &\textit{an undercompressive shock if } \mu_j(u_1) \in \big[\mu_j^{\star\star}(u_0), \mu_j^\star(u_0)\big). \end{aligned} \tag{2.25a}$$

In the second case the shock also satisfies the (stronger) Liu criterion.

(2) *Suppose that* (2.10b) *holds. A shock connecting* u_0 *to* $u_1 = w_j(\mu_j(u_1); u_0)$ *is*

$$\begin{aligned} &\textit{a Lax shock if } \mu_j(u_1) \geq \mu_j(u_0) \ \textit{ or } \ \mu_j(u_1) \leq \mu_j^{\star\star}(u_0); \\ &\textit{a rarefaction shock if } \mu_j(u_1) \in \big(\mu_j^\star(u_0), \mu_j(u_0)\big); \\ &\textit{an undercompressive shock if } \mu_j(u_1) \in \big(\mu_j^{\star\star}(u_0), \mu_j^\star(u_0)\big]. \end{aligned} \tag{2.25b}$$

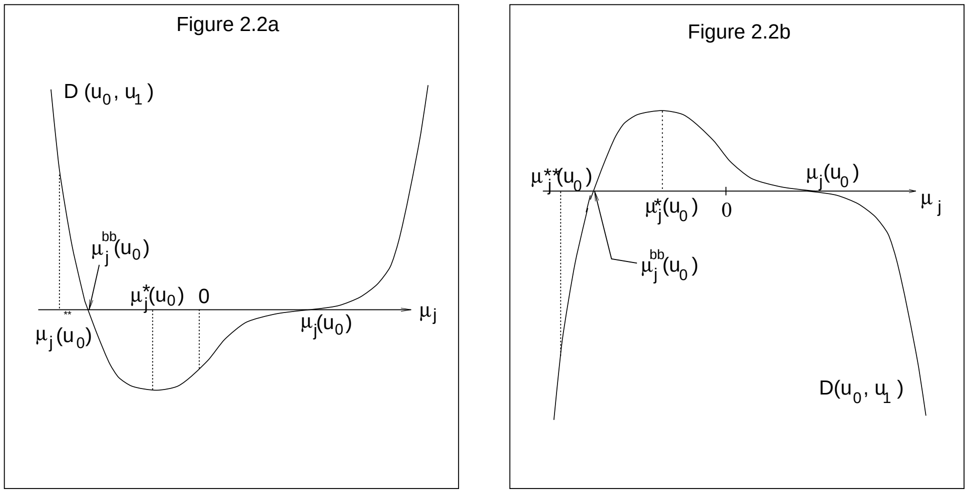


FIG. 2.2. *Entropy dissipation for* (a) *the case* (2.10a), (b) *the case* (2.10b).

In the first case the shock also satisfies the (stronger) Liu criterion.

LEMMA 2.5. *Let* u_0 *be given with* $\mu_j(u_0) \geq 0$ *and consider the Hugoniot curve* $\mathcal{H}_j(u_0)$. *Suppose that* (2.10a) *(resp.,* (2.10b)*) holds.*

(1) *The entropy dissipation* $\mu_j \to D(u_0, w_j(\mu_j; u_0))$ *vanishes at* $\mu_j(u_0)$ *and at a point* $\mu_j^{\flat\flat}(u_0)$ *in the interval* $\big(\mu_j^{\star\star}(u_0), \mu_j^\star(u_0)\big)$. *The entropy dissipation is decreasing (resp., increasing) for* $\mu_j < \mu_j^\star(u_0)$, *increasing (resp., decreasing) for* $\mu_j > \mu_j^\star(u_0)$, *and achieves a negative maximum value (resp., a positive maximum value) at the critical point of the wave speed, that is,* $\mu_j^\star(u_0)$.

(2) *A shock satisfying* (2.8) *cannot be a rarefaction shock. As a corollary, a nonclassical shock is undercompressive and satisfies* $\mu_j \in \big(\mu_j^{\flat\flat}(u_0), \mu_j^\star(u_0)\big)$ *(resp.,* $\mu_j \in \big(\mu_j^{\star\star}(u_0), \mu_j^{\flat\flat}(u_0)\big)$*).*

(3) *Any shock satisfying the Liu criterion* (2.18) *also satisfies the entropy inequality* (2.8).

For u_l and u_r given in $\mathcal{U}$, the Riemann problem (2.1), (2.11) admits up to a P-parameter family of solutions containing N separated wave fans, each of them being composed of (at most) two waves. Specifically we obtain the following description of the classical and nonclassical waves.

Consider a j-wave fan with left-hand state u_0 and right-hand state u with $\mu_j(u_0) \geq 0$. For $j \notin \mathbf{P}$, the wave fan is either a rarefaction wave if $\mu_j(u) > \mu_j(u_0)$, or a classical shock if $\mu_j(u) < \mu_j(u_0)$. For $j \in \mathbf{P}$, we have the following.

Case A. Assume that (2.10a) holds and $j \in \mathbf{P}$. Assume first that $\mu_j(u_0) > 0$. The j-wave fan using only classical waves contains

(1) either a rarefaction from u_0 to $u \in \mathcal{O}_j(u_0)$ if $\mu_j(u) > \mu_j(u_0)$,

(2) a classical shock from u_0 to $u \in \mathcal{H}_j(u_0)$ if $\mu_j(u) \in \big(\mu_j^\star(u_0), \mu_j(u_0)\big)$,

(3) or a classical shock from u_0 to $u^\star := w_j(\mu^\star(u_0); u_0\big)$ followed by an attached rarefaction connecting to $u \in \mathcal{O}_j(u^\star)$ if $\mu_j(u) < \mu_j^\star(u_0)$.

This completes the description of the classical wave curve $\mathcal{W}_j^c(u_0)$ for Case A.

THEOREM 2.6A. *The* j*-wave fan may also contain a nonclassical* j*-shock connecting* u_0 *to* any state $u^\flat \in \mathcal{H}_j(u_0)$ *with* $\mu_j(u^\flat) \in \big(\mu_j^{\flat\flat}(u_0), \mu_j^\star(u_0)\big)$ *followed by*

(1) *either a* nonattached *rarefaction connecting* $u^\flat$ *to* $u \in \mathcal{O}_j(u^\flat)$ *if* $\mu_j(u) < \mu_j(u^\flat)$,

(2) *or by a classical shock connecting $u^\flat$ to $u \in \mathcal{H}_j(u^\flat)$ if $\mu_j(u) > \mu_j(u^\flat)$.*

This defines a two-parameter family of u that can be reached from u_0 by nonclassical solutions. For a given $u^\flat$, the classical shock with largest strength and connecting $u^\flat$ to some $u = u^\sharp \in \mathcal{H}_j(u^\flat)$ is characterized by the condition $\bar\lambda_j(u^\flat, u^\sharp) = \bar\lambda_j(u_0, u^\flat)$ and, in that situation, one also has $u^\sharp \in \mathcal{H}_j(u_0)$. In particular the nonclassical shock with largest possible strength connects the point $u^{\flat\flat} := w_j(\mu_j^{\flat\flat}(u_0); u_0)$ to the point $u^{\sharp\sharp} := w_j(\mu^{\sharp\sharp}(u_0); u^{\flat\flat})$, where $\mu^{\sharp\sharp}(u_0)$ is defined by $u^{\sharp\sharp} \in H_j(u_0)$. Moreover one has

$$\mu_j^{\star\star}(u_0) \le \mu_j^{\flat\flat}(u_0) \le \mu_j^{\flat}(u_0) \le \mu_j^{\star}(u_0) \le \mu_j^{\sharp}(u_0) \le \mu_j^{\sharp\sharp}(u_0) \le \mu_j(u_0). \tag{2.26}$$

In the special case that $\mu_j(u_0) = 0$, the j-wave curve is the j-integral curve issuing from u_0.

Case B. Assume that (2.10b) holds and $j \in \mathbf{P}$. Assume first that $\mu_j(u_0) > 0$. The j-wave fan using only classical waves contains

(1) either a classical shock connecting u_0 to $u \in \mathcal{H}_j(u_0)$ if either $\mu_j(u) \ge \mu_j(u_0)$ or $\mu_j(u) \le \mu_j^{\star\star}(u_0)$,

(2) a rarefaction connecting u_0 to $u \in \mathcal{O}_j(u_0)$ if $\mu_j(u) \in \big[0, \mu_j(u_0)\big]$,

(3) or a rarefaction wave connecting u_0 to a point u_1, followed by an attached classical shock connecting to $u \in \mathcal{H}_j(u_1)$ with $\mu_j(u) = \mu_j^{\star\star}(u_1)$, if $\mu_j(u) \in \big(\mu_j^{\star\star}(u_0), 0\big)$. (In this case the set of u does not describe a rarefaction or shock curve.)

This completes the description of the classical wave curve $\mathcal{W}_j^c(u_0)$.

THEOREM 2.6B. *The j-wave fan may also contain*

(1) *either a rarefaction to $u \in \mathcal{O}_j(u_0)$ if $\mu_j(u) \in \big(0, \mu_j(u_0)\big)$, possibly followed by a* nonattached *nonclassical shock connecting u_1 to u, if $\mu_j(u) \in \big(\mu_j^{\star\star}(u_1), \mu_j^{\flat\flat}(u_1)\big)$ (in this case the set of u does not describe a rarefaction or shock curve),*

(2) *or a classical shock to $u_1 \in \mathcal{H}_j(u_0)$ with $\mu_j(u_1) > \mu_j(u_0)$, followed by a nonclassical shock connecting to $u \in \mathcal{H}_j(u_1)$, if $\mu_j(u) \in \big(\mu_j^{\star\star}(u_1), \mu_j^{\flat\flat}(u_1)\big)$.*

This defines a two-parameter family of u that can be reached from u_0 by nonclassical solutions.

Assume finally that $\mu_j(u_0) = 0$. Then the j-wave curve is the j-Hugoniot curve issuing from u_0 and correspond to classical shocks.

Based on these results, we introduce the following terminology. Given u_0, the set of all states that can be reached using only j-waves will be called the *j-wave set* issuing from u_0 and be denoted by $\mathcal{S}_j(u_0)$ by analogy with the notion of j-wave curve known for classical solutions. We shall call a curve in the wave set a *composite curve* when it is not a part of a rarefaction or shock curve. The wave set in both cases (2.10a) and (2.10b) is represented in Figures 2.3(a) and 2.3(b), respectively. The case that $\mu_j(u_0) < 0$ is analogous and is omitted. We now give a proof of Lemmas 2.4 and 2.5 and Theorem 2.6.

REMARK 2.7. (1) Our analysis shows that, under the assumptions made in this section, the Lax inequalities and the Liu criterion are *equivalent* (Lemma 2.4), which, at first, may appear surprising. The Lax inequalities are sufficient to select a unique solution for shocks with small amplitude near a point where $\nabla\lambda_j \cdot r_j$ vanishes. The Liu criterion is necessary for shocks of moderate amplitude when the product $\nabla\lambda_i \cdot r_i$ changes sign several times along the Hugoniot curve.

(2) When the system (2.1) has a sufficiently large family of entropies (e.g., when $N \le 2$), the formulas (2.28)–(2.29) derived below may be used to establish the converse of item (3) of Lemma 2.5, i.e., limits of regularizations compatible with all entropies (such as (2.2) with $D_\epsilon = \epsilon\, \partial_x u_\epsilon$), necessarily satisfy the Liu criterion.

(3) It may be of interest to search for the weakest constraint on undercompressive

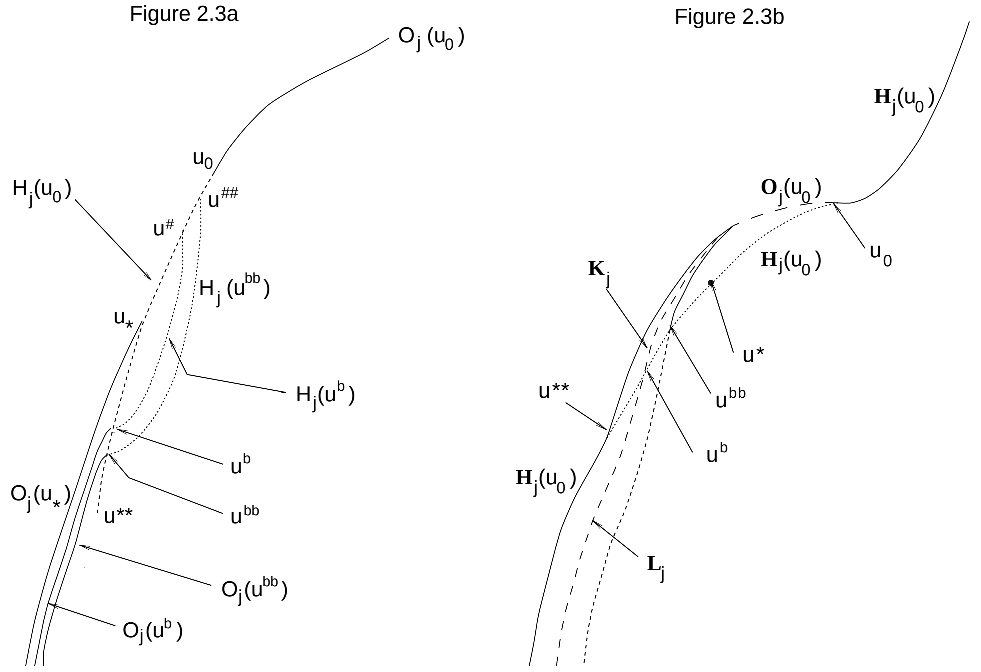


FIG. 2.3. *Wave set* $\mathcal{S}_j(u_0)$ *issuing from* u_0 *for* (a) *the case* (2.10a), (b) *the case* (2.10b).

shocks that can result from imposing one entropy inequality like (2.8). We shall say that a subset $\mathcal{W}_j^{max}(u_0)$ of $\mathcal{U}$ is a *maximal j-wave set* for the system (2.1) if it contains all the j-wave sets for arbitrary entropies. For instance a maximal wave set for the case (2.10a) is obtained by taking $\mu_j^{bb}(u_0) = \mu_j^{\star\star}(u_0)$ in Theorem 2.6; this follows readily from the formula (2.28)–(2.29). In the scalar case with $N = 1$ and $f(u) = u^3$, one has $\mu_j(u_0) = u_0$ and $\mu_j^{\star\star}(u_0) = -2\,u_0$. The scalar case is degenerated and $\mathcal{W}^c(u_0) = \mathcal{W}^{nc}(u_0) = \mathcal{W}^{max}(u_0) = \mathbb{R}$; the interval $[-2\,u_0, u_0]$ is the maximal interval of states that can be reached from u_0 by using a classical or nonclassical shock. □

Proof of Lemma 2.4. Consider for instance the case (2.10a), the case (2.10b) being similar. Lemma 2.3 states that the function $\mu_j \to \bar{\lambda}_j(u_0, w_j(\mu_j; u_0)) - \bar{\lambda}_j(w_j(\mu_j; u_0))$ is positive for $\mu_j > \mu_j^\star(u_0)$ and negative for $\mu_j < \mu_j^\star(u_0)$. On the other hand the function $\mu_j \to \bar{\lambda}_j(u_0, w_j(\mu_j; u_0)) - \lambda_j(u_0)$ is positive for $\mu_j < \mu_j^{\star\star}(u_0)$ or $\mu_j > \mu_j(u_0)$ and negative for $\mu_j \in \big(\mu_j^{\star\star}(u_0), \mu_j(u_0)\big)$. The classification follows easily from these two properties. □

Proof of Lemma 2.5. Using the compatibility condition on the entropy pair, i.e., $\nabla F^T = \nabla U^T Df$, and the Rankine–Hugoniot relation (2.13), the entropy dissipation for a shock connecting u_0 to $w_j(\mu_j; u_0)$ is found to be

$$
\begin{aligned}
&(2.27)\\
&D(u_0, w_j(\mu_j; u_0))\\
&= \int_{\mu_j(u_0)}^{\mu_j} \nabla U(w_j(\zeta_j)) \left\{\bar{\lambda}_j(u_0, w_j(\mu_j)) - Df(w_j(\zeta_j))\right\} \frac{dw_j}{d\zeta_j}(\zeta_j)\, d\zeta_j,\\
&= \int_{\mu_j(u_0)}^{\mu_j} \frac{dw_j}{d\zeta_j}(\zeta_j) \cdot \nabla^2 U(w_j(\zeta_j)) \left\{\bar{\lambda}_j(u_0, w_j(\zeta_j)) \big(w_j(\zeta_j) - u_0\big) - f(w_j(\zeta_j)) + f(u_0)\right\} d\zeta_j.
\end{aligned}
$$

Using once more the Rankine–Hugoniot relation, we get

$$D(u_0, w_j) = \int_{\mu_j(u_0)}^{\mu_j} \left\{ \bar{\lambda}_j(u_0, w_j(\mu_j)) - \bar{\lambda}_j(u_0, w_j(\zeta_j)) \right\} m_j(\zeta_j)\, d\zeta_j, \tag{2.28}$$

where

$$m_j(\zeta_j) := \frac{dw_j}{d\zeta_j}(\zeta_j) \cdot \nabla^2 U(w_j(\zeta_j)) \big(w_j(\zeta_j) - u_0\big) \tag{2.29}$$

has the same sign as $\mu_j - \mu_j(u_0)$. The system (2.1) being strictly hyperbolic, it can be checked that

$$\frac{dw_j}{d\mu_j} \cdot \nabla^2 U(w_j) \;=\; l_j(w_j),$$

which, combined with (2.16ii), shows that $m_j(\zeta_j) > 0$ for $\zeta_j \neq \mu_j(u_0)$.

The occurrence of nonclassical shocks depends on the sign of the entropy dissipation. The integrand in (2.28) has the same sign as $\bar{\lambda}_j(u_0, w_j(\mu_j)) - \bar{\lambda}_j(u_0, w_j(\zeta_j))$, which is nonpositive when the Liu entropy criterion (2.18) holds. It follows that the entropy dissipation is negative as long as the Liu criterion holds. This proves item (3) of Lemma 2.5.

When, instead, the shock satisfies the inequalities (2.21), we have

$$\bar{\lambda}_j(u_0, w_j(\mu_j)) - \bar{\lambda}_j(u_0, w_j(\zeta_j)) \geq 0. \tag{2.30}$$

This indeed is an easy consequence of the facts that $\mu_j \to \bar{\lambda}_j(u_0, w_j(\mu_j))$ is a monotone function (see Lemma 2.3) and that $\bar{\lambda}_j(u_0, u_0) = \lambda_j(u_0) \leq \bar{\lambda}_j(u_0, w_j(\mu_j))$. Combining (2.28) and (2.30) shows that the entropy dissipation is negative for rarefaction shocks. This proves item (2) of Lemma 2.5.

Finally we can establish item (1) by differentiating the formula (2.28) with respect to μ_j:

$$\frac{\partial}{\partial \mu_j} D(u_0, w_j) = \int_{\mu_j(u_0)}^{\mu_j} \frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j)\, m_j(\zeta_j)\, d\zeta_j.$$

This yields a relation between the derivative of the entropy dissipation and that of the shock speed:

$$\frac{\partial}{\partial \mu_j} D(u_0, w_j) = b(w_j) \frac{\partial}{\partial \mu_j} \lambda_j(u_0, w_j), \qquad b(w_j) := \int_{\mu_j(u_0)}^{\mu_j} m_j(\zeta_j)\, d\zeta_j, \tag{2.31}$$

with $C_1\, |w_j - u_0|^2 \leq b(w_j) \leq C_2\, |w_j - u_0|^2$ for some positive constants C_1 and C_2. Note that the dissipation has a critical point either when the shock speed has a critical point or at the point u_0.

From the properties of the shock speed in Lemma 2.3, it follows therefore that $D(u_0, w_j)$ is decreasing for $\mu_j < \mu_j^\star(u_0)$ and increasing for $\mu_j > \mu_j^\star(u_0)$. From its definition, it is clear that $D(u_0, w_j)$ vanishes at $\mu_j(u_0)$. Moreover, we checked that it is positive for $\mu_j < \mu_j^{\star\star}(u_0)$. Therefore there exists a unique point, say, $\mu_j^{\flat\flat}(u_0)$, in the interval $\big(\mu_j^{\star\star}(u_0), \mu_j^\star(u_0)\big)$ where the dissipation vanishes. This completes the proof of Lemma 2.5. □

Proof of Theorem 2.6. We construct the wave set $\mathcal{S}_j(u_0)$ for $u_0 \in \mathcal{U}$ and $j \in \mathbf{P}$. The construction for $j \notin \mathbf{P}$ is classical and $\mathcal{S}_j(u_0)$ is the classical wave curve $\mathcal{W}_j(u_0)$.

Case A. For $u_0 \in \mathcal{M}_j$, either of the conditions (2.8) or (2.18) shows that the wave set $\mathcal{W}_j^{nc}(u_0)$ coincides locally with the integral curve $\mathcal{O}_j(u_0)$. This is because the wave speed is increasing when moving away from u_0 in either direction. The construction is complete for $u_0 \in \mathcal{M}_j$.

We now consider a point u_0 away from the manifold. For definiteness we assume that $\mu_j(u_0) > 0$; the other case could be treated similarly. The construction of the wave curve will use the values $\mu_j^{\star\star}(u_0) < \mu_j^\star(u_0) \le \mu_j(u_0)$ introduced in Lemma 2.3.

For $\mu_j > \mu_j(u_0)$, the state u_0 can be connected to any point on $\mathcal{O}_j(u_0)$ since the wave speed λ_j is increasing for μ_j increasing. Therefore the wave curve $\mathcal{W}_j(u_0)$ coincides with the rarefaction curve $\mathcal{O}_j(u_0)$ for $\mu_j \ge \mu_j(u_0)$.

For μ_j decreasing from $\mu_j(u_0)$, the shock speed is decreasing as long as μ_j remains larger than the critical value $\mu_j^\star(u_0)$. Therefore all the points in the Hugoniot curve $\mathcal{H}_j(u_0)$ with $\mu_j \in [\mu_j^\star(u_0), \mu_j(u_0)]$ can be reached from u_0 by a classical shock satisfying the Liu criterion. According to Lemma 2.5, the entropy dissipation remains negative in the whole range $\mu_j \in [\mu_j^{\flat\flat}(u_0), \mu_j(u_0)]$. Thus the points of the Hugoniot curve $\mathcal{H}_j(u_0)$ with $\mu_j \in [\mu_j^{\flat\flat}(u_0), \mu_j^\star(u_0)]$ can also be reached from u_0 but, now, with a nonclassical shock.

These are the only admissible solutions with a single j-wave issuing from u_0.

Consider now an admissible one-wave solution joining u_0 to u_1. If $\mu_j(u_1) > \mu_j^\star(u_0)$, then no further j-wave can be constructed from u_1. The state $u_1^\star$ with $\mu_j(u_1) = \mu_j^\star(u_0)$ can be connected to any point u_2 in the rarefaction curve $\mathcal{O}_j(u_1)$ with $\mu_j(u_2) \le \mu_j^\star(u_0)$. This covers the whole range of values μ_j and corresponds to the classical wave curve.

We now describe all nonclassical solutions with two j-waves. Consider an admissible one-wave solution joining u_0 to $u^\flat$ with $\mu_j(u^\flat) \in \big[\mu_j^{\flat\flat}(u_0), \mu_j^\star(u_0)\big)$. According to Lemma 2.3, the wave speed is increasing with μ_j decreasing from $\mu_j(u^\flat)$, so $u^\flat$ can be connected to any point u_2 in the rarefaction curve $\mathcal{O}_j(u^\flat)$ with $\mu_j(u_2) \le \mu_j(u^\flat)$. Observe that the nonclassical shock is not attached to the rarefaction fan, i.e.,

$$\bar{\lambda}_j(u_0, u^\flat) < \lambda_j(u^\flat). \tag{2.32}$$

This describes all the solutions containing a nonclassical shock followed by a rarefaction; no further j-wave may follow the rarefaction.

Consider again an admissible one-wave solution joining u_0 to $u^\flat$ with $\mu_j(u^\flat) \in \big[\mu_j^{\flat\flat}(u_0), \mu_j^\star(u_0)\big)$. By (2.32), the shocks with small strength issuing from $u^\flat$ have a larger speed than that of the nonclassical shock, i.e., $\bar{\lambda}_j(u^\flat, u_2) \approx \lambda_j(u^\flat) > \bar{\lambda}_j(u_0, u^\flat)$ for all states u_2 close to $u^\flat$. Hence the speeds have the proper ordering and $u^\flat$ may be connected to any $u_2 \in \mathcal{H}_j(u^\flat)$, at least in the small. Such a shock is also admissible (according to the Liu criterion) since the wave speed is decreasing when μ_j increases (Lemma 2.3.).

This construction can be continued, for $u^\flat$ fixed, until u_2 violates either of the two conditions

$$\bar{\lambda}_j(u^\flat, u_2) > \bar{\lambda}_j(u_0, u^\flat) \text{ or} \tag{2.33}$$

$$D(u^\flat, u_2) \le 0. \tag{2.34}$$

Actually, as $\mu_j(u_2)$ increases from $\mu_j(u^\flat)$, one reaches a maximum value $\mu_j^\sharp$, in which equality holds in (2.33), while the shock is still classical and therefore (2.34) still holds.

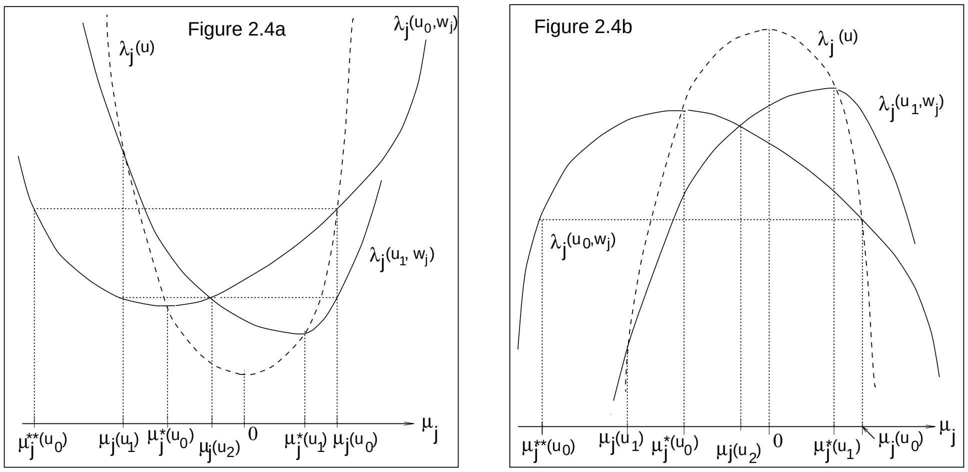


Fig. 2.4. *Graphs of the two shock speeds for* (a) *the case* (2.10a), (b) *the case* (2.10b).

To check the latter, consider the graphs of the two functions $h(\mu_j) := \bar{\lambda}_j(u_0, w_j(\mu_j; u_0))$ and $k(\mu_j) := \bar{\lambda}_j(u^\flat, w_j(\mu_j; u^\flat))$. See Figure 2.4(a). By symmetry of the Rankine–Hugoniot relation, one has $\bar{\lambda}_j(u^\flat, u_0) = \bar{\lambda}_j(u_0, u^\flat)$, so

$$s := h(\mu_j(u^\flat)) = k(\mu_j(u_0)). \tag{2.35}$$

In view of their monotonicity properties, the two graphs must intersect at exactly one point $\mu_j^\sharp$ in the interval $\big(\mu_j(u^\flat), \mu_j(u_0)\big)$. We define $u_2^\sharp$ by the conditions $\mu_j(u_2^\sharp) = \mu_j^\sharp$ and $u_2^\sharp \in \mathcal{H}_j(u^f lat)$.

We claim that, actually,

$$h(\mu_j^\sharp) = k(\mu_j^\sharp) = s \qquad \text{and} \qquad u_2^\sharp \in \mathcal{H}_j(u_0). \tag{2.36}$$

Namely, from the Rankine–Hugoniot relations

$$-s\big(u^\flat - u_0\big) + f(u^\flat) - f(u_0) \;=\; 0 \qquad \text{and} \qquad -\,s\big(u_2^\sharp - u^\flat\big) + f(u_2^\sharp) - f(u^\flat) \;=\; 0,$$

we deduce that $-s\big(u_2^\sharp - u_0\big) + f(u_2^\sharp) - f(u_0) \;=\; 0$, which proves (2.36).

It follows (see Figure 2.4(a)) that (2.33) holds for all $\mu_j(u_2) < \mu_j(u_2^\sharp)$, and the equality holds in (2.33) at the critical value $u_2^\sharp$. Moreover, since $\mu_j(u_2^\sharp) < \mu_j^\star(u_0)$, the shock speed is decreasing on the interval $\big(\mu_j(u^\flat), \mu_j(u_2^\sharp)\big)$ and any shock from $u^\flat$ to u_2 (with $\mu_j(u_2) \le \mu_j(u_2^\sharp)$) satisfies the Liu criterion.

We have the inequalities $\mu_j(u^\flat) < \mu_j^\star(u_0) < \mu_j(u_2^\sharp) < \mu_j(u_0)$. As $\mu_j(u^\flat)$ increases, $\mu_j(u_2^\sharp)$ decreases and eventually both quantities approach the limiting value $\mu_j^\star(u_0)$. As $\mu_j(u^\flat)$ decreases, $\mu_j(u_2^\sharp)$ increases and eventually $\mu_j(u^\flat)$ approaches the limiting value $\mu_j^{\flat\flat}(u_0)$, while $\mu_j(u_2^\sharp)$ approaches some limiting value, say, $\mu_j^{\sharp\sharp}(u_0)$. It is tedious but straightforward to check from the properties of the wave speeds that no third wave can follow a two-wave fan. See Figure 2.2(a) for a representation of the wave set $\mathcal{S}_j(u_0)$.

Case B. For $u_0 \in \mathcal{M}_j$, it is not hard to see, using either of the conditions (2.8) or (2.18), that $\mathcal{W}_j(u_0)$ coincides locally with the Hugoniot curve $\mathcal{H}_j(u_0)$. This is because the wave speed is decreasing when moving away from u_0 in either direction. The construction is complete for $u_0 \in \mathcal{M}_j$.

Consider the case $\mu_j(u_0) > 0$. For $\mu_j > \mu_j(u_0)$, the state u_0 can be connected to any point on $\mathcal{H}_j(u_0)$ since the wave speed is decreasing for μ_j increasing. For $\mu_j < \mu_j(u_0)$, the wave speed is, locally, increasing for μ_j decreasing. So u_0 can be connected to a point on $\mathcal{O}_j(u_0)$ by a rarefaction. This remains possible until μ_j reaches the value 0. It is also possible to connect any point $u_1 \in \mathcal{O}_j(u_0)$ with $\mu_j(u_1) \in \big[0, \mu_j(u_0)\big]$ to a point $u_2 \in \mathcal{H}_j(u_1)$ provided

$$\bar{\lambda}_j(u_1, u_2) = \lambda_j(u_1). \tag{2.37}$$

This construction covers the range $\mu_j \in \big[\mu_j^{\star\star}(u_0), 0\big]$. It is also possible to connect u_0 directly to a point $u \in \mathcal{H}_j(u_0)$ with $\mu_j(u) \leq \mu_j^{\star\star}(u_0)$, since the shock speed in this range satisfies the Liu criterion.

This completes the construction of the classical wave curve $\mathcal{W}_j^c(u_0)$.

We now describe all nonclassical solutions with two j-waves. Consider an admissible one-wave solution from u_0 to u_1. Suppose first $\mu_j(u) \in \big(0, \mu_j(u_0)\big)$ so that $u_1 \in \mathcal{O}_j(u_0)$. One can connect u_1 to $u_2 \in \mathcal{H}_j(u_0)$ by a shock provided both conditions

$$\bar{\lambda}_j(u_1, u_2) \geq \lambda(u_1), \tag{2.38}$$

$$D(u_1, u_2) \, \leq \, 0 \tag{2.39}$$

hold. From the graph of the entropy dissipation, we know that (2.39) is equivalent to

$$\mu_j(u_2) \leq \mu_j^{\flat\flat}(u_0).$$

In view of the graph of the shock speed, (2.38) reads

$$\mu_j^{\flat\flat}(u_1) \leq \mu_j(u_2) \leq \mu_j^{\star}(u_1).$$

Since we always have $\mu_j^{\flat\flat}(u_0) \in \big[\mu_j^{\star\star}(u_0), \mu_j^{\star}(u_0)\big]$, it follows that the admissible interval in the case under consideration is $\mu_j(u_2) \in \big[\mu_j^{\star\star}(u_0), \mu_j^{\flat\flat}(u_0)\big]$. Moreover such a shock is classical only when $\mu_j(u_2) \leq \mu_j^{\star\star}(u_0)$, that is, only when $\mu_j(u_2) = \mu_j^{\star\star}(u_0)$.

Suppose now that $\mu_j(u) \geq \mu_j(u_0)$ so that $u_1 \in \mathcal{H}_j(u_0)$. One can connect u_1 to a point $u_2 \in \mathcal{H}_j(u_1)$ provided

$$\bar{\lambda}_j(u_1, u_2) \geq \bar{\lambda}_j(u_0, u_1) \tag{2.40}$$

and

$$D(u_1, u_2) \, \leq \, 0. \tag{2.41}$$

The condition (2.41) is equivalent to saying $\mu_j(u_2) \leq \mu_j^{\flat\flat}(u_1)$. As μ_j decreases from $\mu_j^{\flat\flat}(u_0)$, the speed $\bar{\lambda}_j(u_1, u_2)$ satisfies (2.40), decreases, and eventually reaches the value $\bar{\lambda}_j(u_0, u_1)$. Since $u_1 \in \mathcal{H}_j(u_0)$ and $u_2 \in \mathcal{H}(u_1)$, the same argument as in the case (2.10a) shows that for that value of μ_j, one has $u_2 \in \mathcal{H}_j(u_0)$. This completes the proof of Theorem 2.6. □

2.3. Selection by kinetic relations. In view of Theorem 2.6, the wave set $\mathcal{S}_j(u_0)$ is a two-dimensional manifold when $j \in \mathbf{P}$. It is our objective now to select a nonclassical wave curve $\mathcal{W}_j^{nc}(u_0)$ in the wave set. Heuristically, it is sufficient to determine one free parameter needed for each nongenuinely nonlinear wave family. One may postulate that for each state u_0, there exists a *single* right state u_1 that can be reached by a nonclassical shock for any $j \in \mathbf{P}$. This is indeed what happens when defining nonclassical shocks as limits of diffusive-dispersive regularizations. We propose to select the admissible nonclassical shocks by considering their entropy dissipation and stipulate the knowledge of an additional jump-like relation on the nonclassical discontinuities. The derivation of such an additional relation for limits of diffusive-dispersive regularizations is discussed later in this subsection.

The following definition stipulates that the entropy dissipation

$$D(u_0, u_1) = -s\ (U(u_1) - U(u_0)) + F(u_1) - F(u_0)$$

of a nonclassical shock, with speed $s = \bar{\lambda}_j(u_0, u_1)$ and connecting u_0 to u_1, is a given "constitutive function" representing certain small-scale properties that have been neglected at the hyperbolic level of modeling. In the following we suppose, for the sake of definiteness, that the condition (2.10a) is satisfied. Dealing with the case (2.10b) requires some modification of the analysis in this subsection. (See also section 3 in which both cases arise.)

We denote by $BV \cap L^\infty$ the space of measurable and bounded functions that have bounded variation in space and time. This space is natural for systems of conservation laws. Functions in $BV \cap L^\infty$ admit traces in a measure theoretic sense [14], so that (2.42) below has a meaning almost everywhere with respect to the one-dimensional Hausdorff measure.

Definition 2.8. *For each $j \in \mathbf{P}$, let $\phi_j : \mathcal{U} \to \mathbb{R}_-$ be a given function. A solution $u(x,t) \in BV \cap L^\infty$ to* (2.1), (2.8) *is called an admissible nonclassical entropy solution if it satisfies the entropy inequality and the entropy dissipation of any nonclassical j-shock in u ($j \in \mathbf{P}$), connecting u_0 to u_1, satisfies*

$$D(u_0, u_1) = \phi_j(u_0). \tag{2.42}$$

We refer to (2.42) as a *kinetic relation* and to ϕ_j as the *kinetic function* for the family j since they determine the propagation of the nonclassical shocks. The kinetic function could also be expressed as a function of the right state u_1 (which need not be equivalent to (2.42)) or—and this is physically more realistic—as a function of a variable "symmetric" in u_0 and u_1, such as the shock speed, or—for problems in fluid dynamics and material science—the mass flux across the discontinuity, etc. Here we shall focus attention on kinetic functions depending solely on the shock speed s, i.e.,

$$D(u_0, u_1) = \varphi(s), \tag{2.43}$$

where φ need be defined only on the union of intervals $\Lambda = \bigcup_{j \in \mathbf{P}} \left[\lambda_j^{\min}, \lambda_j^{\max}\right]$. For scalar conservation laws and under suitable monotonicity conditions, the kinetic function can always be expressed as a function of the shock speed. The same is true for the kinetics generated by diffusive-dispersive regularizations for the systems of two equations studied later in sections 3–5.

In many physical systems, the entropy dissipation is related to the mechanical energy and may be viewed as *a force driving the propagation* of the nonclassical shocks; it is natural to provide a one-to-one relationship between the propagation speed and the driving force. This standpoint was emphasized by Abeyaratne and Knowles [1] for propagating phase boundaries in solids undergoing phase transformations.

In the following we show that the kinetic relation selects a unique curve $\mathcal{W}_j^{nc}(u_0)$ corresponding to nonclassical solutions in the wave set $\mathcal{S}_j(u_0)$. Denote by $D_j^\star(u_0)$ the *maximal negative value* of the entropy dissipation $D(u_0, u_1)$ along the Hugoniot curve $\mathcal{H}_j(u_0)$:

$$D_j^\star(u_0) = \min_{u_1 \in \mathcal{H}_j(u_0)} D(u_0, u_1).$$

Actually the maximum is achieved at the critical value $\mu_j^\star(u_0)$ for the shock speed. Consider also the entropy dissipation as a function of s, say, $d^\star(s)$ defined as

$$d^\star(s) = \max \big\{ D_j^\star(u_0) \,|\, u_0 \in \mathcal{U},\, j \in \mathbf{P}, \quad \lambda_j\big(u_0, w_j(\mu_j^\star(u_0); u_0)\big) = s \big\}. \tag{2.44}$$

(The value is taken to be $-\infty$ when no u_0 satisfies the constraint.) Note that $D_j^\star$ and $d^\star$ are computable from the expression of the flux f in the examples studied in sections 3–5 below.

Theorem 2.9 below shows that knowing the entropy dissipation of the admissible nonclassical shocks determines a unique solution of the Riemann problem. To solve the Riemann problem, we assume that $\big\{ r_k(u, u') \big\}$ is a basis of $\mathbb{R}^N$ for arbitrary u, $u' \in \mathcal{U}$. (This is always true when R is small enough.)

Theorem 2.9. *Suppose that the system satisfies the condition* (2.10a).

(1) *For $j \in \mathbf{P}$, let $\phi_j : \mathcal{U} \to \mathbb{R}_-$ be a continuous function satisfying*

$$D_j^\star(u_0) \le \phi_j(u_0) \le 0 \qquad \textit{for all } u_0 \in \mathcal{U},\, j \in \mathbf{P}. \tag{2.45}$$

Let $u_0 \in \mathcal{U}$ and $j \in \mathbf{P}$ be given. From the wave set $\mathcal{S}_j(u_0)$, there exists a unique wave curve $\mathcal{W}_j^{nc}(u_0)$ using nonclassical shocks satisfying the kinetic relation (2.42). *For u_l and u_r in $\mathcal{U}$, the Riemann problem* (2.1), (2.11) *admits a unique solution in the class of admissible nonclassical entropy solutions obtained by intersection of the curves $\mathcal{W}_j^{nc}$. Furthermore, the solution depends continuously in the L^1 norm upon its end states.*

(2) *Let $\varphi : \Lambda \to \mathbb{R}_-$ be a Lipschitz continuous function satisfying*

$$d^\star(s) \le \varphi(s) \le 0, \quad \frac{d\varphi}{ds}(s) \le 0 \qquad \textit{for all } s \in \Lambda. \tag{2.46}$$

For the kinetic relation (2.43), *the conclusions are the same as in Case* 1.

(3) *In both Cases* 1 *and* 2 *above, there exist two values $\mu_j^\flat(u_0)$ and $\mu_j^\sharp(u_0)$, with*

$$\mu_j^{\star\star}(u_0) \le \mu_j^{\flat\flat}(u_0) \le \mu_j^\flat(u_0) \le \mu_j^\star(u_0) \le \mu_j^\sharp(u_0) \le \mu_j^{\sharp\sharp}(u_0) \le \mu_j(u_0), \tag{2.47}$$

such that the nonclassical wave curve is composed of the following four pieces:

$$\mathcal{W}_j^{nc}(u_0) = \begin{cases} \mathcal{O}_j(u_0) & \textit{for all } \mu_j \ge \mu_j(u_0), \\ \mathcal{H}_j(u_0) & \textit{for all } \mu_j^\sharp(u_0) \le \mu_j \le \mu_j(u_0), \\ \mathcal{H}_j(u^\flat) & \textit{for all } \mu_j^\flat(u_0) \le \mu_j < \mu_j^\sharp(u_0), \\ \mathcal{O}_j(u^\flat) & \textit{for all } \mu_j \le \mu_j^\flat(u_0), \end{cases}$$

where $u^\flat := w_j\big(\mu_j^\flat(u_0); u_0\big)$. The curve $\mathcal{W}_j^{nc}(u_0)$ is continuous and monotone in the parameter μ_j, and is of class C^2 except at $\mu_j = \mu_j^\sharp(u_0)$, where it is generally only Lipschitz continuous.

We can recover the classical curve $\mathcal{W}_j^c(u_0)$ with the (maximal) choice

$$\phi_j(u_0) = D_j^\star(u_0). \tag{2.48}$$

In that case the classical and nonclassical shocks in the solution have the same propagation speed, and the two waves are indistinguishable in the (x,t) plane. On the other hand it is not possible to use part of the classical wave curve, say, for values $\mu_j > \mu_j^c$, and switch to the nonclassical wave curve, say, for values $\mu_j < \mu_j^c$, at least as far as a Riemann solution depending continuously upon its end states is sought. The latter seems to be a natural requirement, at least in view of the examples studied so far in the literature. Furthermore the classical wave curve $\mathcal{W}_j^c(u_0)$ is always admissible, since Definition 2.8 does not prevent us from solving the Riemann problem by using classical waves only. Therefore, even after imposing the kinetic relation, there exist *two* wave curves to choose from for each nongenuinely nonlinear family, $\mathcal{W}_j^c(u_0)$ and $\mathcal{W}_j^{nc}(u_0)$. It would be interesting to connect this nonuniqueness with instability in solutions to an augmented diffusive-dispersive system with vanishing small-scale parameters.

Proof of Theorem 2.9. Let $u_0 \in \mathcal{U}$ and $j \in \mathbf{P}$ be given. In view of the definition (2.44) of the maximal entropy dissipation and the assumption (2.45), the criterion (2.42) selects a unique nonclassical shock along the Hugoniot curve $\mathcal{H}_j(u_0)$, say, $u^\flat = w_j(\mu_j^\flat(u_0), u_0)$. Once this state is selected, the construction in Theorem 2.6 determines a unique wave curve $\mathcal{W}_j^{nc}(u_0)$ having the form described in item (3) of the theorem. This curve is continuous in the parameter μ_j which by construction is monotone increasing along it. It is of class C^2 at the point $\mu_j(u_0)$ and $\mu_j^\flat(u_0)$ since classical rarefaction curves and shock curves have second-order contact. Finally, along the wave curve, the speeds of the (rarefaction or shock) waves change continuously. To see that, at the point $\mu_j^\sharp(u_0)$, one has to compare, on one hand, the shock speed of the nonclassical shock and, on the other hand, the shock speeds of the nonclassical shock and the classical one. Actually all three terms coincide at $\mu_j^\sharp(u_0)$:

$$\lim_{\substack{\mu_j \to \mu_j^\sharp(u_0) \\ \mu_j > \mu_j^\sharp(u_0)}} \bar{\lambda}_j(u_0, w_j(\mu_j, u_0)) = \lim_{\substack{\mu_j \to \mu_j^\sharp(u_0) \\ \mu_j < \mu_j^\sharp(u_0)}} \bar{\lambda}_j\big(u^\flat, w_j(\mu_j; u^\flat)\big) = \bar{\lambda}_j\big(u_0, u^\flat\big).$$

The continuous dependence of the wave speeds implies the L^1 continuous dependence of the solution. Finally, having constructed the Lipschitz continuous wave curves $\mathcal{W}_j^{nc}$ for $j \in \mathbf{P}$ and the smooth wave curves $\mathcal{W}_j^c$ for $j \notin \mathbf{P}$, and using the condition that $\big\{r_k(u,u')\big\}$ is a basis of $\mathbb{R}^N$ for arbitrary u, u', we can solve the Riemann problem with data in $\mathcal{U}$: combining together the wave curves, we apply the theorem of implicit functions for Lipschitz continuous curves. (For a reference see Isaacson and Temple [29].) The Riemann problem admits a unique solution, at least with data in $B(u_*, R') \subset \mathcal{U}$, with $R' << R$. This proves the items (1) and (3).

In order to use the criterion (2.43), one observes that the entropy dissipation $D(u_0,u_1)$ along the Hugoniot curve—when expressed as a function of the shock speed s—is increasing from its lower value $D_j^\star(u_0)$ at $s^\star = \bar{\lambda}_j\big(u_0, w_j(\mu_j^\star(u_0); u_0)\big)$ to the value 0 at $s = \bar{\lambda}_j\big(u_0, w_j(\mu_j^{\flat\flat}(u_0); u_0)\big)$. On the other hand, the function $\varphi(s)$ is assumed to be decreasing in the same interval and by (2.44), (2.46), one has $\varphi(s^\star) \geq d^\star(s) \geq D_j^\star(u_0)$. Thus there exists a unique point $\mu_j = \mu_j^\flat(u_0)$ such that the kinetic relation (2.43) is satisfied. This wave curve shares the same properties as that in the case (2.42). □

Remark 2.10. (1) The assumption that the kinetic function be a decreasing function of the shock speed may be motivated in the following way. Consider a scalar conservation law ($N = 1$) with the flux $f(u) = u^3$ and the entropy $U(u) = u^2/2$. Consider a *linear* relation for nonclassical shocks, say, between the left state u_0 and the right state u_1,

$$u_1 = g(u_0) := \beta\, u_0. \tag{2.49}$$

According to the theory in this section, one must have $\beta \in (-1, -1/2)$. Plugging (2.49) into the definition of the entropy dissipation $D(u_0, u_1)$ the kinetic relation corresponding to (2.49) can be computed:

$$\begin{aligned}\varphi(s) := D\big(u_0, g(u_0)\big) &= -\,(1+\beta)(1-\beta)^3\, u^4\\ &= -\,(1+\beta)(1-\beta)^3(1+\beta+\beta^2)^{-2}\, s^2,\end{aligned}$$

which indeed is a decreasing function of s in the interesting range $s > 0$.

(2) In the classical solution, the value of the intermediate state (if any) in the Riemann solution varies continuously as $u_1 \in \mathcal{W}_j^c(u_0)$ describes the wave curve; the solution in the (x,t) plane varies continuously in the L^1 norm and its total variation is a continuous function of the end points. For the nonclassical wave curve, the wave speeds only are continuous, and the total variation of the Riemann solution is not a continuous function of the endpoints. □

To conclude this section, we explain how to determine the kinetic function, needed in (2.42) or (2.43). Consider a sequence of solutions u^ϵ to a regularized version of (2.1) of the form (2.2). Assume for the sake of this presentation that the u^ϵ remain bounded in the total variation norm and converge to a limiting solution u to (2.1), (2.5). Suppose also that the system admits an entropy pair that is compatible with the regularization (2.2). We know that the entropy inequality (2.5) is too lax to guarantee uniqueness for the Riemann problem. Another Rankine–Hugoniot relation, in addition to the set of conservation laws contained in (2.1), is in principle sufficient to select a unique nonclassical solution.

The concepts of entropy and entropy dissipation are fundamental in the theory of hyperbolic conservation laws. It seems mathematically natural to go beyond the entropy *inequality* (2.8) and instead write the entropy *balance*:

$$\partial_t U(u) + \partial_x F(u) = \mu_U \le 0. \tag{2.50}$$

Here μ_U is a bounded, nonpositive Borel measure, which provides partial information on the small-scale effects in the regularization sequence that generated the solution u. The dissipation measure generated by a regularization (2.2) satisfying the condition (2.3) is

$$\mu_U := w - \star \lim_{\epsilon \to 0}\ \epsilon\, \partial_x u_\epsilon^T \nabla^2 U(u_\epsilon) B_1(u_\epsilon) \partial_x u_\epsilon. \tag{2.51}$$

Since u solves (2.1), the measure μ_U has its support included in the union of the set of points of approximate discontinuity of u.

The mass of the measure along the curve of discontinuity is the entropy dissipation $D(.,.)$.

Of course the knowledge of the measure μ_U in (2.50) is required only for nonclassical shocks, since the propagation of a *classical* shock is uniquely determined by the Rankine–Hugoniot relation

$$-\bar\lambda_j(u_0, u_1)\,\big(u_1 - u_0\big) + f(u_1) - f(u_0) = 0$$

and the entropy *inequality*

$$-\bar\lambda_j(u_0,u_1)\ (U(u_1)-U(u_0))+F(u_1)-F(u_0)\le 0.$$

The entropy dissipation measure μ_U for a *nonclassical* shock, as determined by (2.51), in general, will depend upon the left state u_0 and the shock speed, $s=\bar\lambda_j(u_0,u_1)$. The kinetic relation generated by (2.2) can be determined, at least at a formal level, from an analysis of admissible traveling wave solutions to (2.2). Different approximations to (2.1) will result, in general, in different kinetic relations. Consider a traveling wave solution $u_\epsilon(x,t)=w((x-s\,t)/\epsilon)$ to (2.2), that is a solution to the ordinary differential equation in $\xi=(x-s\,t)/\epsilon$

$$-s\,w'+f(w)'=\big(B_1(w)\,w'\big)'+\big(B_2(w)\,w''\big)' \tag{2.52}$$

satisfying the following boundary conditions

$$\begin{aligned}&\lim_{\xi\to-\infty}w(\xi)=u_0, &&\lim_{\xi\to\infty}w(\xi)=u_1,\\ &\lim_{\xi\to\pm\infty}w'(\xi)=0, &&\lim_{\xi\to\pm\infty}w''(\xi)=0.\end{aligned} \tag{2.53}$$

The equation (2.52) can be integrated once:

$$-s\,(w-u_0)+f(w)-f(u_0)=B_1(w)\,w'+B_2(w)\,w''. \tag{2.54}$$

The internal structure of the nonclassical shock is represented by the trajectory $\xi\to w(\xi)$, which can be used to determine the entropy dissipation measure. Namely, at the hyperbolic level we have

$$\begin{aligned}D(u_0,u_1)&=-\bar\lambda_j(u_0,u_1)\,\big(U(u_1)-U(u_0)\big)-F(u_1)+F(u_0)\\ &=\int_{\mathbb R}\nabla U(w)\cdot\big(-\bar\lambda_j(u_0,u_1)+Df(w)\big)\,w'\,d\xi\\ &=-\int_{\mathbb R}w'\cdot\nabla^2U(w)\cdot\big(-\bar\lambda_j(u_0,u_1)\,(w-u_0)+f(w)-f(u_0)\big)\,d\xi.\end{aligned}$$

Using (2.54) for the traveling wave and the conditions (2.3), we obtain

$$D(u_0,u_1)=-\int_{\mathbb R}(w')^T\nabla^2U(w)B_1(w)\,w'\,d\xi\le 0. \tag{2.55}$$

In the examples arising in continuum mechanics, at least, the entropy dissipation for a nonclassical shock, computed from (2.55), can be expressed as a function of the state u_0 (or, equivalently, u_1). (See also section 4.1.)

3. Nonclassical shocks in elastodynamics (1).

We now turn to a model arising in the theory of elastic materials, which is strictly hyperbolic and admits two nongenuinely nonlinear characteristic fields. This section restricts attention to the Riemann problem and extends the analysis of section 2 to arbitrarily large initial data.

3.1. Preliminaries.

Consider the system of elastodynamics

$$\begin{aligned}\partial_t v-\partial_x\sigma(w)&=0,\\ \partial_t w-\partial_x v&=0,\end{aligned} \tag{3.1}$$

where the real-valued functions v and w represent the velocity and gradient deformation, respectively. The stress-strain law is assumed to have the form

$$\sigma(w) = w^3 + m^2\, w, \qquad m > 0. \tag{3.2}$$

The focus here is on Riemann data

$$v(x,0),\, w(x,0) = \begin{cases} v_l, w_l, & x < 0, \\ v_r, w_r, & x > 0, \end{cases} \tag{3.3}$$

for constants $v_l, w_l, \dots$. We note that (3.1)–(3.2) is invariant under any of the transformations:

$$w \to -w, \quad v \to -v, \tag{3.4i}$$

$$v \to v + \bar{v} \quad \text{(for any constant } \bar{v}\text{)}, \tag{3.4ii}$$

$$x \to -x, \quad v \to -v. \tag{3.4iii}$$

We may write (3.1) in the general form (2.1) by setting $u = (v,w)$, $f(u) = -\big(\sigma(w), v\big)$. The system is strictly hyperbolic with eigenvalues $\lambda_1(v,w) = -c(w) < 0 < \lambda_2(v,w) = c(w)$, where the sound speed is defined by $c(w) = \sqrt{3\,w^2 + m^2}$. Since the wave speeds are independent of v, the notation $\lambda_1(w) = -c(w)$ and $\lambda_2(w) = c(w)$ is also used. The wave speeds are strictly separated: they keep different signs and are bounded away from zero. The right eigenvectors may be chosen as $r_i(v,w) = (\pm c(w), 1)$ for $i = 1, 2$.

We consider the wave curves for the system (3.1). The Hugoniot locus $\mathcal{H}_1(v_0, w_0)$ consists of all the states (v_1, w_1) connected to (v_0, w_0) on the left by a discontinuity with speed $s < 0$. Similarly, $\mathcal{H}_2(v_0, w_0)$ corresponds to the discontinuities with speed $s > 0$. The Rankine–Hugoniot condition gives

$$-s = \frac{v - v_0}{w - w_0} = \frac{\sigma(w) - \sigma(w_0)}{v - v_0}. \tag{3.5}$$

A discontinuity connecting (v_0, w_0) to (v, w) therefore travels with speed $s = \pm\, \bar{c}(w_0; w)$, where we use the notation $\bar{c}(w_0; w) = \sqrt{w_0^2 + w_0\, w + w^2 + m^2}$. Observe that $\bar{c}(w; w) = c(w)$. We emphasize that $\bar{c}(w_0; w)$ is the magnitude of the shock speed and is always positive. From (3.5) we obtain

$$\mathcal{H}_1(v_0, w_0) \;=\; \Big\{ v \in \mathbb{R} \,|\, v - v_0 \;=\; \bar{c}(w_0; w)\,(w - w_0) \Big\}, \tag{3.6}$$

$$\mathcal{H}_2(v_0, w_0) \;=\; \Big\{ v \in \mathbb{R} \,|\, v - v_0 \;=\; -\,\bar{c}(w_0; w)\,(w - w_0) \Big\}. \tag{3.7}$$

In addition, the rarefaction waves are based on the integral curves of the vector fields r_j:

$$\mathcal{O}_1(v_0, w_0) \;=\; \Big\{ v \in \mathbb{R} \,|\, v - v_0 \;=\; \int_{w_0}^{w} c(z)\, dz \Big\}, \tag{3.8}$$

$$\mathcal{O}_2(v_0, w_0) \;=\; \Big\{ v \in \mathbb{R} \,|\, v - v_0 \;=\; -\int_{w_0}^{w} c(z)\, dz \Big\}. \tag{3.9}$$

The system (3.1)–(3.2) is not genuinely nonlinear since $\nabla\lambda_1(w)\cdot r_1(w) = -3\,w/c(w)$ and $\nabla\lambda_2(w)\cdot r_2(w) = 3\,w/c(w)$, which vanish on the (one-dimensional) manifold $\mathcal{M} = \mathcal{M}_1 = \mathcal{M}_2 = \big\{(v,w)\,|\, w = 0\big\}$. In order to uniquely solve the Riemann problem, we now apply appropriate entropy criteria. Away from the line $w = 0$, the system has two genuinely nonlinear fields; therefore, for shocks with small amplitude, the Lax shock inequalities may be used.

3.2. Liu's construction of a unique solution. Here we briefly summarize the Liu's construction for the system (3.1). For a point (v, w) in $\mathcal{H}_1(v_0, w_0)$, the Liu entropy criterion implies the Lax shock inequalities, $-c(w_0) \geq -\bar{c}(w_0; w) \geq -c(w)$, and, as pointed out in section 2, is actually equivalent to them since the stress-strain relation has a single inflexion point. Defining

$$\kappa = w/w_0, \tag{3.10}$$

and using the expressions for $c(w)$ and $\bar{c}(w_0; w)$, one sees that the admissible region for $\mathcal{H}_1(v_0, w_0)$ consists of all (v, w) with

$$\kappa \in (-\infty, -2] \cup [1, +\infty)\,. \tag{3.11}$$

For $\mathcal{H}_2(v_0, w_0)$, the shock speed is positive and the Liu criterion leads to the interval

$$\kappa \in [-1/2, 1]\,. \tag{3.12}$$

Note in passing that the intervals found in (3.11) and (3.12) are independent of m. We now utilize (3.11)–(3.12) and construct the classical wave curves $\mathcal{W}_j^c(v_0, w_0)$. Consider a point (v_0, w_0) with $w_0 > 0$. By (3.4ii), $\mathcal{W}_j^c(v_0', w_0)$ for $v_0' \neq v_0$ is a suitable translate of $\mathcal{W}_j^c(v_0, w_0)$, while (3.4i) allows the construction for $w_0 > 0$ to be simply extended to the case $w_0 < 0$.

The wave curves are easily defined locally. These curves are $\mathcal{H}_1(v_0, w_0)$, $\mathcal{O}_1(v_0, w_0)$, $\mathcal{H}_2(v_0, w_0)$, and $\mathcal{O}_2(v_0, w_0)$ for values $w > w_0$, $w < w_0$, $w < w_0$, and $w > w_0$, respectively. Note that since $\nabla\lambda_i \cdot r_i = \pm 3\,w/c(w)$ changes signs only along curves crossing $w = 0$, we see immediately that the curves $\mathcal{H}_1(v_0, w_0)$ and $\mathcal{O}_2(v_0, w_0)$ may be extended to all points (v, w) such that $w > w_0$. These two curves correspond to functions $w \to v(w)$ that are increasing and decreasing, respectively, according to the formulas (3.6) and (3.9).

We now turn to those wave curves which cross the line $w = 0$. For $0 < w \leq w_0$, we have $\nabla\lambda_i \cdot r_i < 0$, so that all points (v, w) in this region, lying on $\mathcal{O}_1(v_0, w_0)$, may be arrived at by a single 1-rarefaction. This construction changes for $w < 0$: when $-2\,w_0 < w < 0$, there is a critical point on the rarefaction curve, say, $(v_*, w_*) \in \mathcal{O}_1(v_0, w_0)$ with $w_* > 0$, for which $\bar{c}(w_0; , w_*) = c(w_*)$. This point satisfies $w_* = -w/2$.

According to the Liu criterion, in order to reach a point (v, w) from (v_0, w_0), having $-2\,w_0 < w < 0$, the solution proceeds along $\mathcal{O}_1(v_0, w_0)$ until it reaches (v_*, w_*), at which point it jumps on $\mathcal{H}_1(v_*, w_*)$ to (v, w). We denote this composite curve by

$$\begin{aligned} \mathcal{K}_1(v_0, w_0) = \Big\{ (v, w) \,|\, &\text{ there exists } (v_*, w_*) \in \mathcal{O}_1(v_0, w_0), \quad 0 < w_* < w_0, \\ &\text{such that } w = -2\,w_* \text{ and } (v, w) \in \mathcal{H}_1(v_*, w_*) \Big\}. \end{aligned} \tag{3.13}$$

It may be shown that along $\mathcal{K}_1(v_0, w_0)$, v is monotone increasing with w. When $w \le -2\,w_0$, the curve $\mathcal{K}_1(v_0, w_0)$ may be continued, by virtue of (3.11), as a single 1-shock, i.e., $(v, w) \in \mathcal{H}_1(v_0, w_0)$, when $w \le -2\,w_0$ and v is thus given by the Rankine–Hugoniot relation (3.5). Note that $\mathcal{K}_1(v_0, w_0)$ joins $\mathcal{H}_1(v_0, w_0)$ at the point $(v_{**}, w_*) = (v_{**}, -2\,w_0) \in \mathcal{H}_1(v_0, w_0)$, and $\mathcal{K}_1(v_0, w_0)$ joins $\mathcal{O}_1(v_0, w_0)$ at the point $(0, v_0) \in \mathcal{O}_1(v_0, w_0)$. We summarize in the next lemma.

LEMMA 3.1. *The classical* 1*-wave curve from a point* $(v_0, w_0), w_0 > 0$*, is the union of four pieces:*

$$\mathcal{W}_1^c(v_0, w_0) \;=\; \begin{cases} \mathcal{H}_1(v_0, w_0) & \textit{for } \; w > w_0, \\ \mathcal{O}_1(v_0, w_0) & \textit{for } \; 0 \le w \le w_0, \\ \mathcal{K}_1(v_0, w_0) & \textit{for } \; -2\,w_0 \le w < 0, \\ \mathcal{H}_1(v_0, w_0) & \textit{for } \; w < -2\,w_0. \end{cases}$$

It is a monotone increasing curve of class $\mathcal{C}^\infty$*, extending from* $(v, w) = (-\infty, -\infty)$ *to* $(v, w) = (+\infty, +\infty)$*.*

The construction of the 2-wave curve is similar and we summarize its properties as follows.

LEMMA 3.2. *The classical* 2*-wave curve from* (v_0, w_0)*, with* $w_0 > 0$*, is the union of three pieces:*

$$\mathcal{W}_2^c(v_0, w_0) \;=\; \begin{cases} \mathcal{O}_2(v_0, w_0) & \textit{for } \; w > w_0, \\ \mathcal{H}_2(v_0, w_0) & \textit{for } \; -w_0/2 \le w \le w_0, \\ \mathcal{O}_2(v_*, w_*) & \textit{for } \; w < -w_0/2, \end{cases}$$

where $(v_*, w_*) \in \mathcal{H}_2(v_0, w_0)$ *and* $w_* = -w_0/2$*. It is a monotone decreasing curve of class* $\mathcal{C}^\infty$*, extending from* $(v, w) = (+\infty, -\infty)$ *to* $(v, w) = (-\infty, +\infty)$*.*

The infinite extent in v of the 2-wave curve follows from the fact that the integral curves in (3.9) have no horizontal asymptotes. This completes the construction of the wave curves based on the Liu criterion. A unique solution exists for arbitrary Riemann data. It can be checked that this solution depends continuously upon its initial states.

3.3. Two-parameter family of nonclassical entropy solutions. We apply Definition 2.1 to the system (3.1) and construct a two-parameter family of solutions. Definition 2.1 is based on a specific convex entropy pair, which we take here to be

$$U(v, w) = \frac{v^2}{2} + \frac{w^4}{4} + m^2\,\frac{w^2}{2}, \qquad F(v, w) = -v\,\sigma(w). \tag{3.14}$$

This choice is based on the physically motivated regularization studied in section 4. A brief computation leads to the following formula for the entropy dissipation:

$$D(v_-, w_-; v_+, w_+) = -s\big(\bar w (m^2 + \bar{w^2})\,[w] + \bar v\,[v]\big)\big(m^2 \bar w + \bar{w^3}\big)\,[v] - \bar v\,[\sigma(w)] \tag{3.15}$$

with $[\alpha] = \alpha_+ - \alpha_-$ and $\bar\alpha = (\alpha_+ + \alpha_-)/2$. We now substitute the Rankine–Hugoniot relations (3.5) to get

$$D(v_-, w_-; v_+, w_+) = \bar w\,\bar{w^2}\;[v] - \bar{w^3}\;[v] = -\frac{1}{2}\bar w\;[w]^2\;[v]\,.$$

The entropy inequality (2.8), (3.14) therefore reduces to

$$\bar w\;[v] \ge 0 \tag{3.16}$$

(for $[w] \neq 0$). If we now utilize (3.6)–(3.7) for $\mathcal{H}_1(v_-, w_-)$ and $\mathcal{H}_2(v_-, w_-)$, we find that

$$D(v_-, w_-; v_+, w_+) = \frac{s}{2}\,[w]^3\,\bar{w}. \tag{3.17}$$

We recall that $s < 0$ for $\mathcal{H}_1(v_-, w_-)$ and $s > 0$ for $\mathcal{H}_2(v_-, w_-)$. From now on we express the entropy dissipation as a function of w_- and w_+ alone: $D(w_-; w_+)$. The admissible nonclassical shocks from (v_-, w_-) to (v_+, w_+) must therefore satisfy

(3.18)
$$|w_+| \geq |w_-| \quad \text{along} \quad \mathcal{H}_1(v_-, w_-), \qquad |w_+| \leq |w_-| \quad \text{along} \quad \mathcal{H}_2(v_-, w_-).$$

Since $\mathcal{H}_1(v_-, w_-)$, restricted by the condition (3.8), forms a nonconnected set, we denote the portion of $\mathcal{H}_1(v_-, w_-)$ with $w_+ \geq w_-$ by $\mathcal{H}_1^+(v_-, w_-)$, while that portion having $w_+ \leq -w_-$ will be denoted by $\mathcal{H}_1^-(v_-, w_-)$.

We now introduce solutions containing nonclassical shocks. Consider a point (v_0, w_0) with $w_0 > 0$. Owing to transformations (3.4), a translation in v_0 simply effects the same translation in the entire solution; furthermore, we can obtain the wave curves for $w_0 < 0$ by switching the signs of both w and v. We begin by discussing the 1-wave curves. As in the classical case, the solution may leave (v_0, w_0) along $\mathcal{O}_1(v_0, w_0)$ and proceed until it reaches the point $(\tilde{v}, \tilde{w})$ with $\tilde{w} = 0$.

Lemma 3.3. *From any point $(v_1, w_1) \in \mathcal{O}_1(v_0, w_0)$, with $0 < w_1 < w_0$, it is possible to jump to a point $(v_2, w_2) \in \mathcal{H}_1^-(v_1, w_1)$ with $w_2 \in [-2\,w_1, -w_1]$.*

Proof of Lemma 3.3. By (3.18), one has $w_2^2 - w_1^2 \geq 0$. In addition, for the shock to follow the rarefaction, one needs $0 > -\bar{c}(w_1; w_2) \geq \lambda_1(w_1)$, so that $(w_2 + 2\,w_1)(w_2 - w_1) \leq 0$. The intersection of these two regions is the interval $-2\,w_1 \leq w_2 \leq -w_1$. Of the points w_2 in this interval, only the right-hand boundary $w_2 = -2\,w_1$ corresponds to a classical shock. ☐

As (v_1, w_1) varies from (v_0, w_0) to $(\tilde{v}, 0)$, along $\mathcal{O}_1(v_0, w_0)$, the set of image points, $\{(v_2, w_2)\}$, of these nonclassical shocks covers a bounded region. We refer to these wave fans as $\mathcal{O}_1$-$\mathcal{H}_1^-$ nonclassical solutions. From (3.18), it is also possible to leave (v_0, w_0) by a shock, i.e., to jump to $(v_1, w_1) \in \mathcal{H}_1(v_0, w_0)$ for $|w_1| \geq |w_0|$. We note that for $w_1 \geq w_0$ and for $w_1 \leq -2\,w_0$, these are classical shocks. In addition we have the following.

Lemma 3.4. *From a point $(v_1, w_1) \in \mathcal{H}_1^+(v_0, w_0)$, it is possible to jump via a nonclassical shock to $(v_2, w_2) \in \mathcal{H}_1^-(v_1, w_1)$ with $-w_0 - w_1 \leq w_2 \leq -w_1$. The points with $w_2 = -w_0 - w_1$ lie on $\mathcal{H}_1(v_0, w_0)$. The region containing a classical shock along $\mathcal{H}_1^+(v_0, w_0)$, followed by a nonclassical shock along $\mathcal{H}_1^-(v_1, w_1)$, extends indefinitely to the left in w_2, and down in v_2.*

Proof of Lemma 3.4. Once again (3.18) gives $|w_2| \geq |w_1|$, and for the nonclassical shock to follow the classical one, one must also have $-\bar{c}(w_1; w_2) \geq -\bar{c}(w_0; w_1)$. Manipulating the expression for s leads to $|2\,w_2 + w_1| \leq |2\,w_0 + w_1|$, and, using the fact that $0 < w_0 < w_1$, this has the solution $-w_0 - w_1 \leq w_2 \leq w_0$. Combining this with the entropy inequality leads to $-w_0 - w_1 \leq w_2 \leq -w_1$.

For $w_2 = -w_1 - w_0$, one has $-\bar{c}(w_0; w_1) = -\bar{c}(w_0; w_2)$ and, by using the Rankine–Hugoniot condition (3.5), one can show that $(v_2, w_2) \in \mathcal{H}_1^-(v_0, w_0)$. Since $w_2 \leq -2w_0$, the point (v_2, w_2) is in the classical portion of this Hugoniot curve.

According to (3.18), a point $(v_1, w_1) \in \mathcal{H}_1^+(v_0, w_0)$ may have w_1 arbitrarily large and positive, so that $w_2 \leq -w_1$ can be arbitrarily large and negative. A calculation using the Hugoniot curve shows that

$$\begin{aligned} v_2 &= v_0 + \bar{c}(w_0; w_1)\,(w_1 - w_0) + \bar{c}(w_1; w_2)\,(w_2 - w_1) \\ &\le v_0 + \bar{c}(w_0; w_1)\,(w_1 - w_0) - 2\,w_1\,c(w_1), \end{aligned}$$

so that as $w_1 \to +\infty$ with (v_0, w_0) fixed, we have $v_2 \le -w_1^2\,(1 + o(1)) \to -\infty$, so the upper boundary, and hence the entire region, tends to negative infinity, as $w_2 \to -\infty$. There is no horizontal asymptote. ∎

We refer to these 2-wave fans as $\mathcal{H}_1^+$-$\mathcal{H}_1^-$ nonclassical solutions. A similar argument shows that no $\mathcal{O}_1$-$\mathcal{H}_1^-$ or $\mathcal{H}_1^+$-$\mathcal{H}_1^-$ wave fan may be connected to *additional* states by a 1-wave.

We now turn to the 2-wave family, again taking (v_0, w_0) with $w_0 > 0$. In this case, $\lambda_2(w) = \sqrt{3\,w^2 + m^2}$ is increasing with w, so that any point $(v_1, w_1) \in \mathcal{O}_2^+(v_0, w_0)$, i.e., with $w_1 \ge w_0$, may be connected to (v_0, w_0) via a 2-rarefaction. We may not continue from (v_1, w_1) to a point $(v_2, w_2) \in \mathcal{H}_2(v_1, w_1)$, since the entropy inequality, which gives $|w_2| \le |w_1|$, and the proper ordering of wave speeds, which implies $w_2 \ge w_1$, have only the degenerate point $(v_2, w_2) = (v_1, w_1)$ of intersection. If instead we leave (v_0, w_0) via $\mathcal{H}_2(v_0, w_0)$, the entropy inequality permits us to proceed to the left, until we reach $(\tilde{v}, \tilde{w}) \in \mathcal{H}_2(v_0, w_0)$ with $\tilde{w} = -w_0$. Note that this shock is nonclassical for $-w_0 \le w_1 < -w_0/2$.

Lemma 3.5. *For $(v_1, w_1) \in \mathcal{H}_2(v_0, w_0)$ with $-w_0 \le w_1 \le -w_0/2$, it is possible to connect to a point $(v_2, w_2) \in \mathcal{H}_2(v_1, w_1)$ with $w_1 \le w_2 \le -w_0 - w_1$. This part of the curve $\mathcal{H}_2(v_1, w_1)$ extends until it reaches a point $(v_2, w_2) = (v_2, -w_0 - w_1) \in \mathcal{H}_2(v_0, w_0)$.*

Proof of Lemma 3.5. Starting from a point $(v_1, w_1) \in \mathcal{H}_2(v_0, w_0)$, we proceed with a 2-shock on the right, to a point $(v_2, w_2) \in \mathcal{H}_2(v_1, w_1)$. The entropy inequality forces $|w_2| \le |w_1|$. In addition, the requirement that $c(w_1; w_2) \ge c(w_0; w_1) \ge 0$ implies that $|2\,w_2 + w_1| \ge |2\,w_0 + w_1|$. Since $w_0 \ge |w_2|$, we must take $w_1 \le 0$. The condition then becomes $|2\,w_2 - |w_1|| \ge 2\,w_0 - |w_1|$, so that w_2 must also be nonpositive. Some manipulation gives $w_2 \le -w_0 - w_1 \le 0$, so that in combination with (3.18) we have $w_2 \in [w_1, -w_0 - w_1]$, and w_1 has the restriction that $w_1 \le -w_0 - w_1$, so that $w_1 \le -w_0/2$. This leads to $w_1 \in [-w_0, -w_0/2]$. At the right-hand end of this interval, $w_1 = -w_0/2$, the shock is classical. ∎

As w_1 varies about the interval $[-w_0, -w_0/2]$, the set $\{(v_2, w_2)\}$ of image points attainable from (v_0, w_0) by a nonclassical shock, followed by a second shock, fill up a bounded region. This second shock is always a classical one, across which w does not change signs. Points on this second shock may not, therefore, be connected to a further rarefaction or shock wave. We now consider rarefaction waves originating at a point on $\mathcal{H}_2(v_0, w_0)$.

Lemma 3.6. *A point $(v_1, w_1) \in \mathcal{H}_2(v_0, w_0)$, with $-w_0 \le w_1 \le -w_0/2$, may be connected to any point $(v, w) \in \mathcal{O}_2(v_1, w_1)$ having $w \le w_1$.*

Proof of Lemma 3.6. Since $\lambda_2(w)$ is increasing for $|w|$ increasing, if points (v_1, w_1) can be found so that $c(w_1) \ge c(w_0; w_1) \ge 0$, then the rarefaction curves $\mathcal{O}_2(v_1, w_1)$ may be continued indefinitely to the left. The condition on wave speeds reduces to $(2w_1 + w_0)(w_1 - w_0) \ge 0$, so that we must have $w_1 \le -w_0/2$. Thus any (v_1, w_1) with $w_1 \in [-w_0, -w_0/2]$ can serve as the origin of a 2-rarefaction. Note that the classical shock-rarefaction occurs for $w_1 = -w_0/2$. ∎

From (3.8)–(3.9), all of the (classical and nonclassical) integral curves have $v \to +\infty$ as $w \to -\infty$. As w_1 varies from $-w_0/2$ to $-w_0$, the set of points that may be reached by a nonclassical 2-shock, followed by a rarefaction, forms an unbounded

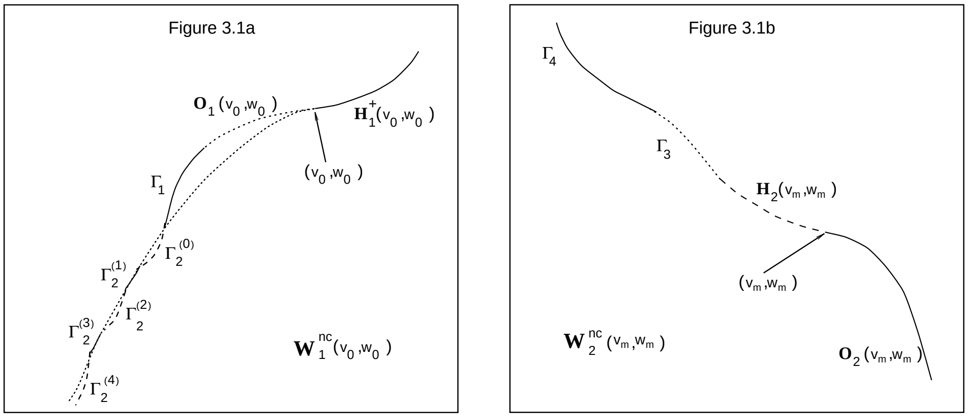


FIG. 3.1. *Wave curve for* (a) *the* 1*-wave family,* (b) *the* 2*-wave family.*

strip in the (v, w)-plane. These rarefaction curves may not be further joined to 2-shock curves, since the entropy inequality and the proper wave speed ordering (the shock must travel faster than the maximum wave speed of the rarefaction fan) lead to incompatible intervals in w. We summarize the above results in this subsection by stating the following.

THEOREM 3.7. *The solutions to* (3.1)–(3.3) *satisfying a single entropy inequality form a one-parameter family in each of the two characteristic fields. The shock speeds* s_1 *and* s_2 *of the nonclassical shocks in the* 1*- and* 2*-wave families, respectively, may be used as the parameters. Given a left-hand state* (v_0, w_0) *and denoting the left-hand state of the nonclassical shock by* (v_-, w_-)*, there are nonclassical solutions in the* 1*-family for* s_1 *satisfying*

$$max\left\{ -\sqrt{w_0^2 + w_0\, w_- + w_-^2 + m^2}\,,\, -\sqrt{3\, w_-^2 + m^2} \right\} \;\le s_1 \;\le\; -\sqrt{w_-^2 + m^2},$$

and in the 2*-family for* s_2 *satisfying*

$$\sqrt{(3/4)w_-^2 + m^2} \;\le\; s_2 \;\le\; \sqrt{w_-^2 + m^2}.$$

3.4. Unique admissible nonclassical entropy solution. In this section, we construct the nonclassical wave curves $\mathcal{W}_j^{nc}(v_0, w_0)$, $j = 1, 2$, displayed in Figure 3.1. For a solution connecting $u_0 = (v_0, w_0)$ to $u_1 = (v_1, w_1)$, we label the successive states, according to increasing wave speed, by $u_0, u_{i_1} = (v_{i_1}, w_{i_1}), u_m = (v_m, w_m), u_{i_2} = (v_{i_2}, w_{i_2})$, and u_1. For classical shocks or rarefactions in the 1-wave and 2-waves curves, the points u_{i_1} and u_{i_2}, respectively, degenerate into u_0 and u_1, respectively. Nonclassical 1-shocks always connect u_{i_1} to a range of u_m, while nonclassical 2-shocks join u_m to a range of u_{i_2}. In the classical cases where shocks are attached to rarefactions, one always has $w_m = -2w_{i_1}$ and/or $w_{i_2} = -w_m/2$.

Depending on u_0 and u_1, a nonclassical shock may appear in either $\mathcal{H}_1$, or $\mathcal{H}_2$, or both. To select the unique nonclassical shock from among the one-parameter families of solutions found in the previous subsection, we will utilize a *kinetic relation*, stipulating that any nonclassical 2-shocks from u_m to u_{i_2} must satisfy

$$D(w_m; w_{i_2}) = \varphi(s), \tag{3.19}$$

where φ is the kinetic function depending upon the shock speed s. For a given left-hand state (v_m, w_m), we show that the kinetic relation produces a unique right-hand state (v_{i_2}, w_{i_2}), where w_{i_2} depends only on w_m, and not on v_m, say, $w_{i_2} = g(w_m)$. From Lemmas 3.3 and 3.4, $w_{i_2} \in [-w_m, -w_m/2)$.

In order to select a unique nonclassical shock in the 1-wave family, a symmetry of system (3.1) is utilized: the nonclassical 1-shock from u_{i_1} to u_m is selected from among the possible nonclassical 1-shocks from u_{i_1}, if the kinetic relation (3.19) is satisfied for $w_{i_2} = w_{i_1}$. We begin with the kinetic relation for $\mathcal{H}_2$-shocks, divorced from their role in the solution of the Riemann problem, and consider nonclassical $\mathcal{H}_2$-shocks from (v_-, w_-) to (v_+, w_+).

Theorem 3.8. *Denote by $I = [m, \infty)$ the range of positive shock speeds, s, and consider a kinetic function $\varphi(s)$ having the following properties:*

$$\frac{d\varphi}{ds} < 0 \quad \textit{for} \quad s \in I, \tag{3.20}$$

$$\varphi(s) \leq 0 \quad \textit{for all} \quad s \in I, \tag{3.21}$$

$$\varphi(s) \geq -\frac{3}{4} s\,(s^2 - m^2)^2 \quad \textit{for } s \in I. \tag{3.22}$$

Then the kinetic relation (3.19) *selects a unique value of the right-hand state $w_+ = g(w_-)$, from among the nonclassical shocks in $\mathcal{H}_2(v_-, w_-)$.*

Proof of Theorem 3.8. Without loss of generality, we take $w_- > 0$. From (3.17), we have

$$D(w_-; w_+) = \bar{c}(w_-; w_+)(w_+ - w_-)^3 (w_+ + w_-)/4, \tag{3.23}$$

and, from the previous subsection, $-w_- \leq w_+ < -w_-/2$ for the $\mathcal{H}_2$ nonclassical shock. The following calculation shows that the entropy dissipation of (3.23) is monotone in w_+:

$$\frac{\partial D(w_-; w_+)}{\partial w_+} = (2\,w_+ + w_-)\,(5\,w_+^2 + 4\,w_+ w_- + 3\,w_-^2 + 4\,m^2)\,\frac{(w_+ + w_-)^2}{2\,\bar{c}(w_-; w_+)}. \tag{3.24}$$

We rewrite the second factor in (3.24) as $3\,w_+^2 + w_-^2 + 2\,(w_+ + w_-)^2 + 4\,m^2 > 0$. So only the first factor may change sign, and therefore

$$\partial D/\partial w_+ < 0 \quad \text{along} \quad \mathcal{H}_2(v_-, w_-) \quad \text{for} \quad w_+ < -w_-/2, \tag{3.25}$$

so that D is monotone decreasing in w_+ for nonclassical 2-shocks.

For fixed $w_- > 0$ and $w_+ < -w_-/2$,

$$\frac{\partial c(w_-; w_+)}{\partial w_+} = \frac{2w_+ + w_-}{2\,\bar{c}(w_-; w_+)} < 0, \tag{3.26}$$

so that combined with (3.25), this shows that in the region of admissible nonclassical 2-shocks, D is increasing with s. Therefore by (3.20), the functions $D(w_-; w_+)$ and $\varphi(s)$ can have at *most* one intersection point.

We now verify that conditions (3.21) and (3.22) ensure one such intersection. Initially, by (3.18), we have $D(w_-; w_+) \leq 0$. Condition (3.21) is therefore a natural upper bound on φ. In addition, the maximum negative entropy dissipation for a given w_- occurs at the "classical" endpoint, $w_+ = -w_-/2$, of the admissible nonclassical interval. At this point, $s = \sqrt{3\,w_-^2/4 + m^2}$ and by (3.24),

$$\begin{aligned} D(w_-; -w_-/2) &= s\,(-3\,w_-/2)^3\,(w_-/2)/4 \\ &= -\frac{3}{4}s\,(s^2 - m^2)^2\,. \end{aligned} \tag{3.27}$$

Thus, if $\varphi(s)$ remains within the bounds (3.21) and (3.22), the kinetic relation (3.19) must have a unique solution. This completes the proof of Theorem 3.8. □

Remark. The construction of Theorem 3.8 cannot be extended to cover the nonclassical 1-shocks, i.e., to the interval $s \in (-\infty, -m]$, as the following argument demonstrates. For the admissible, nonclassical 1-shock region, $-2w_- < w_+ \le -w_-$, one finds again that the entropy dissipation D is monotonically increasing with s. This compels us to take $\varphi'(s) < 0$ in $(-\infty, -m]$. On the other hand, we also require $\varphi(s) \le 0$, and $\varphi(s) \ge 3s(s^2 - m^2)/4$, with this latter function increasing to zero at the right-hand endpoint of the interval. No kinetic function ϕ can possibly satisfy this combination of constraints over the whole interval of s.

We therefore abandon the idea of having independently prescribed kinetics for each of the families of waves. Instead, we will show existence and uniqueness for the Riemann problem under an assumption of *symmetric kinetics.* With symmetric kinetics, *a nonclassical $\mathcal{H}_1$-shock from w_{i_1} to w_m is selected if the kinetic relation for $\mathcal{H}_2$ selects a shock from w_m to $w_{i_2} = w_{i_1}$.* For the case of nonclassical shocks in both families, this assumption results in the two nonclassical shocks being mirror images of each other across the w-axis in the (x, w)-plane. We will see in section 4 that a numerical scheme for (3.1) produces such symmetric nonclassical shocks.

We motivate a symmetric choice of w_{i_1} by noting that system (3.1) is invariant under the transformation $x \to -x, \quad v \to -v$. Thus to any nonclassical 2-shock from (v_m, w_m) to (v_{i_2}, w_{i_2}), there corresponds a nonclassical 1-shock from (v_{i_1}, w_{i_1}) to (v_m, w_m) with $w_{i_1} = w_{i_2}$. These shocks are actually antisymmetric in v and have $v_{i_1} = 3v_{i_2} - 2v_m$. Whether or not such nonclassical shocks are admissible depends on the relative values of w_0 and w_{i_1}, as the following lemma shows.

Lemma 3.9. *Consider a point u_0, where $w_0 > 0$ without loss of generality. For $0 < w_{i_1} < w_0$, the nonclassical* 1*-shock from w_{i_1} to w_m where $w_m = h(w_{i_1})$ is determined by the kinetic relation* (3.19) *is always an admissible nonclassical* 1*-shock. For $w_{i_1} > w_0$, the nonclassical shock from w_{i_1} to w_m, where $w_{i_1} = w_{i_2}$ and $w_{i_2} = g(w_m)$, is determined from the $\mathcal{H}_2$ kinetics, is only admissible if $w_m \in (-w_{i_2} - w_l, -w_{i_2}]$.*

Remark. The function $h(w_{i_1})$ for the symmetric kinetics in the 1-wave family is the inverse of $g(\cdot)$, which yields the right-hand state for nonclassical 2-shocks. Since, as we will show in Theorem 3.10, the function $g(\cdot)$ is monotone in its argument, such an inverse exists and is well defined.

Proof of Lemma 3.9. For $w_{i_1} > 0$, we have $w_m < 0$ and $w_{i_2} > 0$. By Lemma 3.3, $w_{i_2} \in I_2 := (-w_m/2, -w_m]$. For $w_{i_1} < w_0$, which corresponds to the rarefaction/nonclassical 1-shock wave-fan, $w_m \in I_1 = (-2w_{i_1}, -w_{i_1}] = (-2w_{i_2}, -w_{i_2}]$, and therefore $w_{i_2} \in I_2$ iff $w_m \in I_1$.

In the case $w_{i_1} > w_0$, Lemma 3.4 stipulates that there can be a nonclassical shock joining u_{i_1} to u_m, if $w_m \in I_3 = (-w_{i_1} - w_0, -w_{i_1}]$, and by the symmetric kinetics assumption, $I_3 = (-w_{i_2} - w_0, -w_{i_2}]$. Meanwhile, for the nonclassical 2-shock, $w_m \in (-2w_{i_2}, -w_{i_2}]$ which contains the interval I_3, since $w_0 < w_{i_2}$. □

Remark. The intervals I_3 and I_4 are almost identical for $w_0 \approx w_{i_1}$. When $w_{i_1} \gg w_0$, however, the interval I_3 becomes an ever-diminishing fraction of I_4, relegated to the upper end containing the "most" nonclassical shocks. As $w_{i_2} \to \infty$, unless the kinetic relation selects $w_{i_2} = -w_m$, no admissible, symmetric 1-shock can be constructed.

We prepare for the construction of the nonclassical wave curves, with a given kinetics, by proving that the 2-shocks selected by (3.19) have $w_+ = g(w_-)$ monotone decreasing in w_-.

THEOREM 3.10. *For a nonclassical* 2*-shock between* (v_-, w_-) *and* (v_+, w_+), *with* $w_+ = g(w_-)$ *selected by the kinetic relation* (3.19),

$$\frac{d\,g(w_-)}{d\,w_-} < 0 \quad \textit{for} \;\; s \in [m, \infty)\,. \tag{3.28}$$

Proof of Theorem 3.10. In light of the Rankine–Hugoniot condition, we may view the selection of a unique right-hand state, w_+, alternatively as the selection of a unique (nonclassical) shock speed, $s(w_-)$. Thus we may reexpress the kinetic relation (3.19) as

$$\mathcal{D}(w_-; s) = \varphi(s). \tag{3.29}$$

Taking the derivative with respect to w_- in (3.29) gives

$$\frac{\partial \mathcal{D}}{\partial w_-} + \frac{\partial \mathcal{D}}{\partial s}\,\frac{\partial s}{\partial w_-} = \varphi'(s)\,\frac{\partial s}{\partial w_-}. \tag{3.30}$$

Rearrangement of (3.30) leads to

$$\frac{\partial s}{\partial w_-} = \frac{\partial \mathcal{D}/\partial w_-}{\varphi' - \partial \mathcal{D}/\partial s}\,. \tag{3.31}$$

Comparing the functions D and $\mathcal{D}$, we find

$$\frac{\partial D}{\partial w_-} = \frac{\partial \mathcal{D}}{\partial w_-} \quad \text{and} \quad \frac{\partial D}{\partial w_+}\,\frac{\partial w_+}{\partial s} = \frac{\partial \mathcal{D}}{\partial s}. \tag{3.32}$$

For the $\mathcal{H}_2$ nonclassical shocks, we have from Theorem 3.8 that $\partial D/\partial w_+ < 0$ and $\partial w_+/\partial s < 0$, so that by (3.32) we have $\partial \mathcal{D}/\partial s > 0$. In addition, since $\varphi' < 0$ by (3.20), the denominator in (3.31) is always negative. We now use the first equality of (3.32) to compute the sign of the numerator in (3.31).

We regard s as a parameter and compute the derivative of (3.23) with respect to w_-, where we have

$$w_+^2 = -w_-\,w_+ + S - w_-^2 \tag{3.33}$$

from the Rankine–Hugoniot relation; here we have defined $S = s^2 - m^2$. We note that $3\,w_-^2/4 \le S \le w_-^2$. Taking the derivative of (3.33) gives $w'_+ = -(w_+ + 2w_-)/(2w_+ + w_-)$. A straightforward calculation from (3.23) using these quantities results in

$$\frac{\partial D}{\partial w_-} = \frac{-2\,\bar{c}(w_-; w_+)}{2w_+ + w_-}\left[S - \frac{3\,w_-}{2}(w_- + w_+)\right](\,S - 3\,w_-^2\,)\,. \tag{3.34}$$

The first factor is positive, since $\bar{c}(w_-; w_+) > 0$ and $2\,w_+ + w_- < 0$. The second factor in (3.34) is greater than or equal to $S - 3\,w_-^2/4$ and so is also positive. Finally, from the above bounds on S, the third factor in (3.33) is negative. Thus from (3.34) we have $\partial D/\partial w_- < 0$ for Case A. This implies, from (3.31), that $\partial s/\partial w_- > 0$. The result (3.28) then follows from the Rankine–Hugoniot condition. This completes the proof of Theorem 3.10. $\square$

Corollary 3.11. *The nonclassical* 1*-shock between* (v_{i_1}, w_{i_1}) *and* (v_m, w_m) *with* $w_{i_1} = w_{i_2}$ *from symmetric kinetics, where* $w_{i_2} = g(w_m)$*, has the monotonicity property*

$$\frac{d\,h(w_{i_1})}{d\,w_{i_1}} < 0 \quad \textit{for} \quad s \in (-\infty, -m]. \tag{3.35}$$

We now turn to construction of the nonclassical wave curves, beginning with $\mathcal{W}_1^{nc}(v_0, w_0)$. The point $u_0 = (v_0, w_0)$ is arbitrary, but we take $w_0 > 0$ here for definiteness. Just as in the Liu construction, the wave-curve may be extended indefinitely to the right, along the classical portion of the $\mathcal{H}_1(v_0, w_0)$ curve. Similarly, $\mathcal{W}_1^{nc}(v_0, w_0)$ may be continued to the left until it reaches the point $(v, w) = (\tilde{v}, 0)$, with $\tilde{v}$ given by (3.10), along the integral curve $\mathcal{O}_1(v_0, w_0)$.

To extend this 1-wave curve into the region with $w < 0$, we utilize the symmetric kinetics. For $0 \le w_{i_1} < w_0$, it is possible by Lemma 3.3 and Corollary 3.11 to connect (v_0, w_0) to a point $(v_m, h(w_{i_1}))$, with $-2w_0 < h(w_{i_1}) \le 0$, by an $\mathcal{O}_1$-$\mathcal{H}_1^-$ wave fan. The union of these points, as w_{i_1} varies between w_0 and zero, is given by the curve

$$\begin{aligned} \Gamma_1 = \Big\{ &(v_m, w_m) \in \mathcal{H}_1^-(v_{i_1}, w_{i_1}) \mid w_m = h(w_{i_1}) \in (-2\,w_{i_1}, -w_{i_1}], \\ &w_{i_1} \in \mathcal{O}_1(v_0, w_0),\ \ 0 < w_{i_1} < w_0 \Big\}. \end{aligned} \tag{3.36}$$

By the monotonicity property of $w_m = h(w_{i_1})$, the left-hand endpoint of Γ_1, which represents a single nonclassical shock, must be the point $(v, w) = (v_0^*, h(w_0)) \in \mathcal{H}_1^-(v_0, w_0)$, where the value of v_0^* is found from the Hugoniot curve (3.6).

According to Lemma 3.9, when $w_{i_1} > w_0$ there will be a nonclassical $\mathcal{H}_1^+$-$\mathcal{H}_1^-$ wave fan, connecting (v_0, w_0) to $(v_m, h(w_{i_1}))$ iff

$$\psi_h(w_{i_1}; w_0) = h(w_{i_1}) + w_{i_1} + w_0 \ge 0 \tag{3.37}$$

holds, where $h(w_{i_1})$ is the value of w_m selected by the kinetic relation for w_{i_1}. Note that the monotonicity property, from Corollary 3.11, of h does not imply the satisfaction or failure of condition (3.37) and, for a very general kinetic function $\varphi(s)$ in Theorem 3.8, there may be successive intervals in $w_{i_1} > w_0$ where nonclassical shocks are alternately allowed or disallowed.

Since $h(w_{i_1})$ changes smoothly with w_{i_1}, we must have $\psi_h(w_{i_1}; w_0) = 0$ in (3.37) just before it becomes negative, for a slightly larger w_{i_1}. From Lemma 3.4, equality in (3.37) implies that $(v_m, h(w_{i_1}))$ lies on the classical shock curve $\mathcal{H}_1^-(v_0, w_0)$.

We therefore augment the symmetric kinetics for the 1-wave nonclassical shocks by the additional requirement that *if, for a given* w_0*, we have* $\psi_h(w_{i_1}; w_0) < 0$*, at some* $w_m = h(w_{i_1})$*, determined from symmetric kinetics, then the point* $(v_m, w_m) \in \mathcal{W}_1^{nc}(v_0, w_0)$ *is chosen by requiring* $(v_m, w_m) \in \mathcal{H}_1^-(v_0, w_0)$.

By Lemma 3.9, $\psi_h(w_0; w_0) \ge 0$. If there is strict inequality, the nonclassical $\mathcal{H}_1^+$-$\mathcal{H}_1^-$ wave fan will persist until $w_{i_1} = \tilde{w}_1$, where ψ_h switches from positive to negative. Note that $\tilde{w}_1 = w_0$ if $\psi_h(w_0; w_0) = 0$. According to our augmented symmetric kinetics, we continue $\mathcal{W}_1^{nc}(v_0, w_0)$ as a portion of $\mathcal{H}_1^-(v_0, w_0)$ until $w_{i_1} = \tilde{w}_2$, where ψ_h changes from negative to positive. The next segment—to the left of the previous one, as w_m is decreasing with increasing w_{i_1}—of $\mathcal{W}_1^{nc}(v_0, w_0)$ will be a nonclassical one, continuing until $w_{i_1} = \tilde{w}_3$, and so on.

This pattern of alternating classical and nonclassical portions of $\mathcal{W}_1^{nc}(v_0, w_0)$ may continue indefinitely. Regardless of the pattern of classical and nonclassical curves, it follows from Lemmas 3.1 and 3.4 that for (v_m, w_m) on $\mathcal{W}_1^{nc}(v_0, w_0)$, we have $v_m \to -\infty$ as $w_m \to -\infty$.

Let $\{\tilde{w}_k\}$, $\quad k = 0, 1, 2, \dots$, with $\tilde{w}_0 = w_0 \leq \tilde{w}_1 < \tilde{w}_2 < \cdots$, be the set of points where $w_{i_1} \geq w_0$ has $\psi_h(w_{i_1}; w_0) = 0$. From the above argument, $\psi_h(w_{i_1}; w_0) > 0$ for $\tilde{w}_{2k} < w_{i_1} < \tilde{w}_{2k+1}$, while $\psi_h(w_{i_1}; w_0) < 0$ when $\tilde{w}_{2k+1} < w_{i_1} < \tilde{w}_{2k+2}$. We can then describe the portions of allowable nonclassical $\mathcal{H}_1^+$-$\mathcal{H}_1^-$ wave fans by

$$\begin{aligned}\Gamma_2^{(2k)} = \Big\{ &(v_m, w_m) \in \mathcal{H}_1^-(v_{i_1}, w_{i_1}) \mid w_m = h(w_{i_1}), (v_{i_1}, w_{i_1}) \in \mathcal{H}_1^+(v_0, w_0), \\ &\tilde{w}_{2k} \leq w_{i_1} \leq \tilde{w}_{2k+1} \Big\}\end{aligned} \tag{3.38}$$

for $k = 0, 1, 2, \dots$. The right-hand endpoint of $\Gamma_2^{(0)}$ represents a single nonclassical shock from (v_0, w_0) to the point $(v_0^*, h(w_0))$, so that this precisely matches the left-hand endpoint of the curve Γ_1, calculated previously. The curve $\mathcal{W}_1^{nc}(v_0, w_0)$ is therefore continuous at $w_m = h(w_0)$. The left-hand endpoint of $\Gamma_2^{(0)}$, as well as both endpoints of $\Gamma_2^{(2k)}$ for $k > 0$, join continuously to the (classical) Hugoniot curve $\mathcal{H}_1^-(v_0, w_0)$, according to Lemma 3.4. We denote the segments of $\mathcal{H}_1^-(v_0, w_0)$, used in this construction, by

$$\begin{aligned}\Gamma_2^{(2k+1)} = \Big\{ &(v_m, w_m) \in \mathcal{H}_1^-(v_0, w_0), \mid w_m = h(w_{i_1}), (v_{i_1}, w_{i_1}) \in \mathcal{H}_1^+(v_0, w_0), \\ &\tilde{w}_{2k+1} \leq w_{i_1} \leq \tilde{w}_{2k+2} \Big\}\end{aligned} \tag{3.39}$$

for $k = 0, 1, 2, \dots$. The curve $\mathcal{W}_1^{nc}(v_0, w_0)$ is then given by

$$\mathcal{W}_1^{nc}(v_0, w_0) = \begin{cases} \mathcal{H}_1^+(v_0, w_0) & \text{for } w > w_0, \\ \mathcal{O}_1(v_0, w_0) & \text{for } 0 \leq w \leq w_0, \\ \Gamma_1 & \text{for } h(w_0) \leq w < 0, \\ \Gamma_2^{(0)} \cup \Gamma_2^{(1)} \cup \Gamma_2^{(2)} \cup \cdots & \text{for } w < h(w_0), \end{cases} \tag{3.40}$$

where $h(w_0) < 0$ is determined by symmetric kinetics, and w_0 is taken to be positive. Together, the above union of curves stretches continuously from $(v, w) = (-\infty, -\infty)$ to $(v, w) = (\infty, \infty)$.

We complete the discussion of $\mathcal{W}_1^{nc}(v_0, w_0)$ by showing that it increases monotonically in v as a function of w. Since the classical portions of this curve are known from Lemma 3.1 to be monotone increasing in w, it remains to show that the nonclassical segments are also increasing. The next lemma shows that, in fact, the curves Γ_1 and $\Gamma_2^{(2k)}$ are monotone increasing.

LEMMA 3.12. *Suppose* $(v_m, w_m) \in \Gamma_1$ *or* $(v_m, w_m) \in \Gamma_2^{(2k)}$. *Then* v_m *is monotonically increasing with* w_m.

Proof of Lemma 3.12. For the point (w_m, v_m) on Γ_1, one calculates that

$$\frac{dv_m}{dw_{i_1}} = -\frac{\left(\bar{c}(w_{i_1}; w_m) - \bar{c}(w_{i_1})\right)^2}{2\,\bar{c}(w_{i_1}; w_m)} + \frac{dw_m}{dw_{i_1}}\left(c(w_{i_1}; w_m) + \frac{(w_m - w_{i_1})(2\,w_m + w_{i_1})}{2\,c(w_{i_1}; w_m)}\right).$$

The first term is nonpositive, while the coefficient of the dw_m/dw_{i_1} can be shown to be positive. Since we have $dh(w_{i_1})/dw_{i_1} < 0$, from Corollary 3.11, it follows that $dv_m/dw_{i_1} < 0$, and so Γ_1 increases from left to right in w. It can further be shown that for $(v_m, w_m) \in \Gamma_2^{(2k)}$,

$$\frac{dv_m}{dw_{i_1}} = (\bar{c}(w_0; w_{i_1}) - \bar{c}(w_m; w_{i_1})) \left(1 - \frac{\bar{c}^2(w_{i_1})}{\bar{c}(w_0; w_{i_1}) - \bar{c}(w_m; w_{i_1})}\right) + \frac{dw_m}{dv_{i_1}} \left(\bar{c}(w_{i_1}; w_m) + \frac{(2\, w_m + w_{i_1})(w_m - w_{i_1})}{2\, \bar{c}(w_{i_1}; w_m)}\right).$$

Since we have the inequalities $\bar{c}(w_{i_1}) \geq \bar{c}(w_0; w_{i_1}) \geq \bar{c}(w_{i_1}; w_m) \geq 0$, the first term is negative, while the coefficient of dw_m/dv_{i_1} is again positive. Applying Corollary 3.11, regarding the sign of $dh(w_{i_1})/dw_{i_1}$, yields the desired monotonicity for v_m as a function of w_m. □

We now turn to the construction of $\mathcal{W}_2^{nc}(v_m, w_m)$. For this discussion, $u_m = (v_m, w_m)$ is taken to be arbitrary. Alternatively, we can view this point as $u_m \in \mathcal{W}_1^{nc}(v_0, w_0)$ for some $u_0 = (v_0, w_0)$. To be definite, we take $w_m > 0$, but this discussion could be extended to $w_m < 0$ with little complication. We are interested in the set of points $u_1 = (v_1, w_1)$ that can be connected to u_m through either a rarefaction, classical shock, shock-rarefaction, or a pair of shocks with positive wave speeds; in the latter two cases, there will be an intermediate state, $u_{i_2} = (v_{i_2}, w_{i_2})$, between u_m and u_1. A specific kinetic function has been imposed, so that the kinetic relation (3.19) selects a unique value $w_{i_2} = g(w_m)$ from among the possible nonclassical shocks in $\mathcal{H}_2(v_m, w_m)$.

As in the classical case (see Lemma 3.2), when $w > w_m$, this portion of $\mathcal{W}_2^{nc}(v_m, w_m)$ is just $\mathcal{O}_2(v_m, w_m)$. Similarly, when $0 < w < w_m$, we have that this section of $\mathcal{W}_2^{nc}(v_m, w_m)$ matches the classical shock curve $\mathcal{H}_2(v_m, w_m)$. To determine how far this classical shock curve penetrates into the region $w < 0$, however, the specific kinetics must be taken into account, as we do below.

The point $u_{i_2} \in \mathcal{H}_2(v_m, w_m)$, with $-w_m \leq w_{i_2} < -w_m/2$, is the unique right-hand state for the nonclassical shock, determined by the kinetic relation (3.19). From u_{i_2}, the solution may be continued, according to Lemma 3.6, along $\mathcal{O}_2(v_{i_2}, w_{i_2})$ for $w < w_{i_2}$. We denote this portion of $\mathcal{W}_2^{nc}(v_m, w_m)$ by

$$\Gamma_4 = \left\{ (v, w) \in \mathcal{O}_2(v_{i_2}, w_{i_2}) \mid w_{i_2} = g(w_m), \quad -\infty < w \leq w_{i_2} \right\}. \tag{3.41}$$

According to Lemma 3.6, one may also continue from u_{i_2} to $u_1 = (v_1, w_1) \in \mathcal{H}_2(v_{i_2}, w_{i_2})$ for $w_{i_2} \leq w_1 \leq -w_m - w_{i_2}$. This portion of $\mathcal{W}_2^{nc}(v_m, w_m)$ will be labeled by

$$\Gamma_3 = \left\{ (v, w) \in \mathcal{H}_2(v_{i_2}, w_{i_2}) \mid w_{i_2} = g(w_m), \quad w_{i_2} \leq w \leq -w_m - w_{i_2} \right\}. \tag{3.42}$$

For $u_1 \in \mathcal{H}_2(v_{i_2}, w_{i_2})$, with $w_1 = -w_m - w_{i_2}$, we saw in Lemma 3.5 that $u_1 \in \mathcal{H}_2(v_m, w_m)$ as well. We may therefore complete the construction of $\mathcal{W}_2^{nc}(v_m, w_m)$ in a continuous manner by extending the classical portion of $\mathcal{H}_2(v_m, w_m)$ until $w = -w_m - w_{i_2}$. This continuous, nonclassical 2-wave curve is then given by

$$\mathcal{W}_2^{nc}(v_m, w_m) = \begin{cases} \mathcal{O}_2(v_m, w_m) & \text{for } w > w_m, \\ \mathcal{H}_2(v_m, w_m) & \text{for } -w_m - w_{i_2} < w \leq w_m, \\ \Gamma_3 & \text{for } w_{i_2} \leq w \leq -w_m - w_{i_2}, \\ \Gamma_4 & \text{for } w < w_{i_2}, \end{cases} \tag{3.43}$$

where $w_{i_2} = g(w_m)$ comes from the kinetic relation (3.19). We now show that the curve $\mathcal{W}_2^{nc}(u_m)$ of (3.43) has v monotone decreasing with w.

LEMMA 3.13. *The curve $\mathcal{W}_2^{nc}(v_m, w_m)$ defined in* (3.43) *is continuous, with v monotone decreasing in w, from $(v, w) = (-\infty, \infty)$ to $(v, w) = (\infty, -\infty)$. Furthermore, $\mathcal{W}_2^{nc}(v_m, w_m)$ is $\mathcal{C}^{\in}$ except at $w = -w_m - w_{i_2}$, where it is merely continuous.*

Proof of Lemma 3.13. The monotonicity of $\mathcal{W}_2^{nc}(v_m, w_m)$ follows from it being the continuous union of four monotone decreasing curves: Γ_4, which is a portion of $\mathcal{O}_2(v_{i_2}, w_{i_2})$, has v decreasing for increasing w by (3.9). This integral curve naturally joins $\mathcal{H}_2(v_{i_2}, w_{i_2})$ at u_{i_2} with second-order contact, so that Γ_3 and Γ_4, together, form a $\mathcal{C}^{\in}$ curve. By (3.7), we have Γ_3 decreasing as w increases.

From Lemma 3.6, Γ_3 and $\mathcal{H}_2(v_m, w_m)$ meet at $w = -w_m - w_{i_2}$, implying continuity. The remaining portion of $\mathcal{W}_2^{nc}(v_m, w_m)$ is classical, and its continuity and monotonicity follow from Lemma 3.2.

The infinite extent, in v, of $\mathcal{W}_2^{nc}(v_m, w_m)$ follows from the divergence of the integral in (3.9), as $|w| \to \infty$. This proves Lemma 3.13. □

Combining Lemmas 3.12 and 3.13, regarding the infinite extent, continuity, and the respective monotonicities of the nonclassical wave curves, $\mathcal{W}_1^{nc}(v_0, w_0)$ and $\mathcal{W}_2^{nc}(v_m, w_m)$, we have in the following theorem our main result of this section.

THEOREM 3.14. *Given a point (v_0, w_0), the Riemann problem for system* (3.1) *with initial data (u_0, u_1), where $u_1 = (v_1, w_1)$ is an arbitrary point, has a unique solution in the class of nonclassical shocks, given a kinetic function $\varphi(s)$ satisfying assumptions* (3.20)–(3.22), *and assuming augmented symmetric kinetics for the* 1*-wave family.*

4. Nonclassical shocks in elastodynamics (2).

4.1. Convergence result.

For the model of section 3, the convergence of some approximations toward weak solutions is easily established, applying the method of compensated compactness (Murat [49], Tartar [60], DiPerna [13]) as we show in this subsection. With no uniqueness result available for nonclassical solutions, only subsequences of solutions can be shown to converge. It is one of the challenging open problems in this area to extend the kinetic relation, introduced in this paper for traveling waves, to more general solutions. This is because the kinetic relation has been introduced for functions of bounded variation, while the compensated compactness approach provides solutions in a functional space of less regular functions (i.e., L^p).

Consider the augmented version of the elastodynamics system:

$$\begin{aligned} \partial_t v - \partial_x \sigma(w) &= \epsilon\, \partial_{xx} v - \alpha\, \epsilon^2\, \partial_{xxx} w, \\ \partial_t w - \partial_x v &= 0, \end{aligned} \tag{4.1}$$

where ϵ and α are positive constants. Here σ is given by (3.2) as in section 3. Regularization terms as in the right-hand side of (4.1) were first studied by Slemrod (see [59] and Fan and Slemrod [15]) for the case that σ is decreasing in some interval, which models phase transitions in materials or in fluids; therein the dispersion term models the capillarity effect of the fluid. As we can demonstrate numerically, the sign of the dispersion term in (4.1) corresponds to that where nonclassical behavior is observed.

As the coefficients in front of the diffusion and dispersion terms vanish, the solutions to (4.1) converge to a nonclassical solution to the hyperbolic model (3.1). Observe that the presence of the dispersion term in the right-hand side of the first equation in (4.1) (and the absence of diffusion in the second equation) prevents obtaining an L^∞ bound by the theory of invariant regions à la Chuey, Conley, and Smoller [7]. The theorem below uses L^p estimates, instead.

Define the internal energy W by $W'(w) = \sigma(w)$. From (3.2) one gets $W(w) = \big(w^4 + 2\,m^2\,w^2\big)/4$.

THEOREM 4.1. (1) *Let $\big(v^\epsilon, w^\epsilon\big)$, with $\alpha \geq 0$ fixed, be a family of solutions to* (4.1) *assuming at $t = 0$ a Cauchy data $\big(v_0^\epsilon, w_0^\epsilon\big)$ satisfying uniform bounds in ϵ in the following spaces:*

$$v_0^\epsilon \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}), \qquad w_0^\epsilon \in L^1(\mathbb{R}) \cap L^4(\mathbb{R}), \qquad \epsilon^{1/2}\,\partial_x w_0^\epsilon \in L^2(\mathbb{R}). \tag{4.2}$$

Then the sequences v^ϵ and w^ϵ remain uniformly bounded in $L^\infty\big(\mathbb{R}_+, L^2(\mathbb{R})\big)$ and $L^\infty\big(\mathbb{R}_+, L^4(\mathbb{R})\big)$, respectively, and converge almost everywhere to limiting functions v and w, solutions to the hyperbolic system (3.1).

(2) *The entropy pair $(U, F) = \big(v^2/2 + W(w), -v\,\sigma(w)\big)$ is compatible in the sense* (2.3) *with the diffusive-dispersive regularization* (4.1). *Limits of traveling wave solutions to* (4.1), *additionally, satisfy the entropy inequality*

$$\partial_t\left(\frac{v^2}{2} + W(w)\right) - \partial_x\big(v\,\sigma(w)\big) \leq 0. \tag{4.3}$$

We do not expect the entropy inequalities

$$\partial_t U(v, w) + \partial_x F(v, w) \leq 0, \tag{4.4}$$

with $U(v, w) \neq v^2/2 + W(w)$ (up to a linear function of v and w), to hold in general.

Proof of Theorem 4.1. The bounds in L^2 and L^4 follow from the following standard energy estimate. Multiplying the first equation in (4.1) by σ and the second one by v, we arrive at

$$\partial_t\big(W(w)+v^2/2\big)-\partial_x\big(v\,\sigma(w)\big) \;=\; -\epsilon\,|\partial_x v|^2+\epsilon\,\partial_x\big(v\,\partial_x v\big)-\alpha\,\epsilon^2\,\partial_x\big(v\,\partial_{xx} w\big)+\alpha\,\epsilon^2\partial_x v\,\partial_{xx} w.$$

Using the second equation in (4.1), we rewrite $\partial_x v\,\partial_{xx} w = \partial_t w\,\partial_{xx} w = \partial_x\big(\partial_t w\,\partial_x w\big) - \partial_t\big(|\partial_x w|^2/2\big)$. Therefore we obtain the following entropy balance:

$$\begin{aligned}&\partial_t\big(W(w) + v^2/2 + \alpha\,\epsilon^2\,|\partial_x w|^2/2\big) - \partial_x\big(v\,\sigma(w)\big)\\ &\;= -\epsilon\,|\partial_x v|^2 + \epsilon\,\partial_{xx}\big(v^2/2\big) - \alpha\,\epsilon^2\,\partial_x\big(v\,\partial_{xx} w\big) + \alpha\,\epsilon^2\partial_x\big(\partial_x v\,\partial_x w\big).\end{aligned} \tag{4.5}$$

This leads to the following uniform bound:

$$\begin{aligned}&\int_{\mathbb{R}}\big(W(w) + v^2/2 + \alpha\,\epsilon^2\,|\partial_x w|^2/2\big)(T)\,dx + \int_0^T\int_{\mathbb{R}}\epsilon\,|\partial_x v|^2\,dxdt\\ &= \int_{\mathbb{R}}\big(W(w) + v^2/2 + \alpha\,\epsilon^2\,|\partial_x w|^2/2\big)(0)\,dx \leq O(1),\end{aligned} \tag{4.6}$$

where we have used (4.2) and where $O(1)$ denotes a constant independent on ϵ.

Multiply the first equation in (4.1) by $\partial_x w$ and integrate in space and time to write, on one hand,

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}}\big(\partial_x w\,\partial_t v - \partial_x w\,\sigma_w(w)\partial_x w\big)\,dxdt\\ &= \left[\int_{\mathbb{R}}\partial_x w\,v\,dx\right]_0^T - \int_0^T\int_{\mathbb{R}}\partial_{xx} v\,v\,dxdt - \int_0^T\int_{\mathbb{R}}\sigma_w(w)\,|\partial_x w|^2\,dxdt\\ &= \int_{\mathbb{R}}\partial_x w(T)\,v(T)\,dx - \int_{\mathbb{R}}\partial_x w(0)\,v(0)\,dx + \int_0^T\int_{\mathbb{R}}|\partial_x v|^2\,dxdt\\ &\quad - \int_0^T\int_{\mathbb{R}}\sigma_w(w)\,|\partial_x w|^2\,dxdt\end{aligned}$$

and, on the other hand,

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}} \partial_x w \left(\epsilon\, \partial_{xx} v - \alpha\, \epsilon^2\, \partial_{xxx} w\right) dxdt \\
&= \int_0^T \int_{\mathbb{R}} \epsilon\, \partial_x w\, \partial_{tx} w\, dxdt + \alpha\, \epsilon^2 \int_0^T \int_{\mathbb{R}} |\partial_{xx} w|^2\, dxdt \\
&= \left[\epsilon \int_{\mathbb{R}} |\partial_x w|^2/2\, dx\right]_0^T + \alpha\, \epsilon^2 \int_0^T \int_{\mathbb{R}} |\partial_{xx} w|^2\, dxdt.
\end{aligned}$$

Observe that

$$\left| \int_{\mathbb{R}} \partial_x w(T)\, v(T)\, dx \right| \le \epsilon \int_{\mathbb{R}} |\partial_x w(T)|^2/2\, dx + \big(2\epsilon\big)^{-1} \int_{\mathbb{R}} |v(T)|^2\, dx,$$

and similarly for the term $\partial_x w(0)\, v(0)$. Finally, combining the above formulas, we obtain

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}} \epsilon\, \sigma_w(w)\, |\partial_x w|^2\, dxdt + \alpha\, \epsilon^2 \int_0^T \int_{\mathbb{R}} |\partial_{xx} w|^2\, dxdt \\
&\quad \le \int_0^T \int_{\mathbb{R}} \epsilon\, |\partial_x v|^2\, dxdt + \epsilon^2 \int_{\mathbb{R}} |\partial_x w(0)|^2\, dx + \int_{\mathbb{R}} |v(T)|^2/2\, dx, + \int_{\mathbb{R}} |v(0)|^2/2\, dx.
\end{aligned} \tag{4.7}$$

Combining (4.6) and (4.7) and using the form (3.2) of the function σ, we obtain the uniform bounds

$$\int_{\mathbb{R}} \big(v(T)^2 + w(T)^2 + w(T)^4\big)\, dx + \int_{\mathbb{R}} \alpha\, \epsilon\, |\partial_x w(T)|^2\, dx \le O(1), \tag{4.8}$$

$$\int_0^T \int_{\mathbb{R}} \big(\epsilon\, |\partial_x v|^2 + \epsilon\, |\partial_x w|^2 + \alpha\, \epsilon^2\, |\partial_{xx} w|^2\big)\, dxdt \le O(1). \tag{4.9}$$

Using the $L^2 \times L^4$ bound derived for the sequence (v_ϵ, w_ϵ), we introduce a Young measure representing possible oscillations in the sequence as $\epsilon \to 0$. The estimates (4.8)–(4.9) are the basis for applying DiPerna's argument in [13], which shows that the Young measure satisfies the so-called Tartar commutation equation. The standard reduction theorem, stating that it must reduce to a Dirac mass, does not apply here since the support of the Young measure is not bounded.

Instead, the work by Shearer [55] and Serre and Shearer [54] based on L^p estimates does apply. The system (3.1) is strictly hyperbolic and the constitutive equation σ possesses a single inflection point. The theorem in [54] implies that there exists a limiting function $(v, w) \in L^\infty(L^2 \times L^4)$ such that the sequence strongly converges to (v, w) in the sense

$$\begin{aligned}
v_\epsilon &\to v \text{ in } L^p \text{ for all } p < 2, \\
w_\epsilon &\to w \text{ in } L^p \text{ for all } p < 2.
\end{aligned} \tag{4.10}$$

Observe that (4.10) suffices for the passage to the limit in (4.1) and in order to derive (3.1): the nonlinearity $\sigma(w)$ is cubic while we have a control of w in L^4 by the entropy estimate (4.6).

Showing that the natural entropy of the system (3.1) is compatible with the regularization (4.1) is easy from (2.3). It is a classical matter (see Schonbek [53] and, also, Hayes and LeFloch [22] for the analogous case of scalar equations with vanishing diffusion and dispersion) to check that, in view of (4.8)–(4.9), the right-hand side of (4.5) converges to zero in the sense of distributions. The entropy flux does converge to its corresponding limit since $\sigma(w^\epsilon)$ converges strongly to $\sigma(w)$. The term $\alpha\,\epsilon\,\partial_t|\partial_x w_\epsilon|^2$ converges to zero in the sense of distributions thanks to (4.8)–(4.9). Let us, equivalently, check that the product of v and $\alpha\,\epsilon^2\,\partial_{xxx}w$ converges to zero. Namely, for each smooth function θ with compact support,

$$\begin{aligned}
&\left|\int_0^T\int_{\mathbb{R}} \epsilon^2\, v\,\partial_{xxx}w\,\theta\,dxdt\right| \\
&\le \left|\int_0^T\int_{\mathbb{R}} \epsilon^2\,\partial_x v\,\partial_{xx}w\,\theta\,dxdt\right| + \left|\int_0^T\int_{\mathbb{R}} \epsilon^2\, v\,\partial_{xx}w\,\partial_x\theta\,dxdt\right| \\
&\le O(1)\,\epsilon^2\,\|\partial_x v\|_{L^2((0,T)\times\mathbb{R})}\,\|\partial_{xx}w\|_{L^2((0,T)\times\mathbb{R})} + O(1)\,\epsilon^2\,\|v\|_{L^2((0,T)\times\mathbb{R})}\,\|\partial_{xx}w\|_{L^2((0,T)\times\mathbb{R})} \\
&\le O(1)\,\epsilon^{1/2} + O(1)\,\epsilon \;\to\; 0.
\end{aligned}$$

Further estimates are needed to treat the entropy term in general, since we know only that $W(w^\epsilon)$ is bounded in $L^\infty_t(L^1_x)$ and, therefore, could a priori converge to a bounded Radon measure. However, in the special case of (smooth) traveling wave solutions to (4.1), it is straightforward to deduce (4.3) follows from the entropy balance (4.5), since all of the terms in the right-hand side of (4.5) have a conservative form but one which is nonpositive. □

We now comment upon the derivation a kinetic relation for (3.1) associated with the regularization (4.1). After rescaling by ϵ, a traveling wave solution (v, w) to (4.1), connecting (v_0, w_0) to (v_1, w_1) and having the speed s, satisfies the following third order system of ODEs:

$$\begin{aligned}
s\,w' + v' &= 0, \\
s\,v' + \sigma(w)' &= -v'' + \alpha\,w'''
\end{aligned}$$

together with the conditions $v(\xi) \to v_0$, $w(\xi) \to w_0$ at $\xi \to -\infty$, and $v(\xi) \to v_1$, $w(\xi) \to w_1$ at $\xi \to +\infty$. We also assume that w', w'', and w''' vanish at $\pm\infty$. Eliminating the variable v, we obtain an equation for the scalar-valued function w:

$$-s^2\,w' + \sigma(w)' = s\,w'' + \alpha\,w'''.$$

Integrating once, we obtain

$$-s^2\,\big(w - w_0\big) + \sigma(w) - \sigma(w_0) = s\,w' + \alpha\,w''. \tag{4.11}$$

Given a value for the shock speed s, there exist up to three states that solve the equation giving the equilibrium points of (4.11), i.e.,

$$-s^2\,\big(w - w_0\big) + \sigma(w) - \sigma(w_0) = 0. \tag{4.12}$$

Namely, these are w_0 itself and (at most) two additional points w_1 and w_2. Since the cubic $\sigma(w) = w^3 + m^2 w$ has no quadratic term, one must have $w_0 + w_1 + w_2 = 0$. Consider the case that w_0 is chosen such that $w_0 > 0$ and $w_1 < w_2 < 0$ which holds in a certain range of values for s.

Consider for instance waves of the second characteristic family propagating with $s > 0$. From the Liu criterion, it follows that a traveling wave connecting w_0 to w_1 represents a classical shock, while a connection from w_0 to w_2 is a nonclassical shock. A typical feature of (4.11) is the following one [62]: there exists a critical value for the slope $s^\sharp$ such that a traveling wave trajectory connecting to w_1 exists for speeds $s > s^\sharp$ and there exists a connection to w_2 when $s = s^\sharp$.

We emphasize that, given w_0, there exist a unique state w_2 and a unique speed such that w_0 and w_2 can be connected by a nonclassical shock. The traveling wave analysis therefore allows us to write, say,

$$w_2 = g(w_0) \qquad \text{and} \qquad s = s(w_0). \tag{4.13}$$

Using (4.13), the entropy dissipation of the nonclassical shocks can be computed as a function of the left state of the shock. This determines the kinetic function ϕ:

$$\phi(w_0) := D(w_0, w_2) = D\big(w_0, g(w_0)\big). \tag{4.14}$$

Provided the relation $s = s(w_0)$ is one-to-one, one can rewrite (4.14) and obtain the kinetic function expressed as a function of the propagation speed, that is,

$$\varphi(s) := \phi(w_0) \qquad \text{with } s^2 = w_0^2 + g(w_0)^2 + w_0\, g(w_0) + m^2. \tag{4.15}$$

The possibility of writing the kinetic function as a function of a single variable (here w), and hence as a function of the speed s, is a special property of the system (3.1) *and* the regularization (4.1). Other regularizations to (3.1), for which a scalar equation like (4.11) could not be derived, may require a kinetic function of the general form $\phi(v_0, w_0)$.

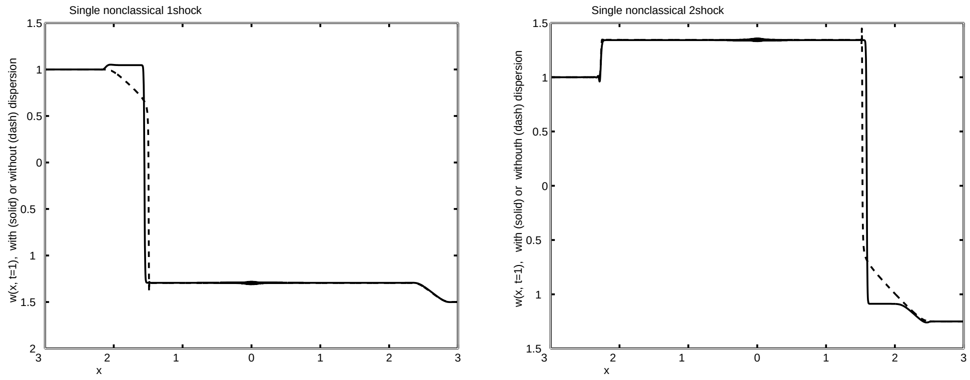
4.2. Numerical experiments. The paper [23] is devoted to the numerical analysis of nonclassical shocks in finite difference schemes. Our purpose here is to illustrate that nonclassical shocks do indeed appear.

In this subsection, we solve the Riemann problem numerically and confirm some of the results enumerated in section 3. We employ the following semidiscrete approximation to the augmented system

$$\begin{aligned} \frac{dv_k}{dt} - \frac{1}{2\,\Delta}\big(\sigma(w_{k+1}) - \sigma(w_{k-1})\big) &= \frac{\epsilon}{\Delta^2}\big(v_{k+1} - 2\,v_k + v_{k-1}\big) \\ &\quad - \frac{\alpha\,\epsilon^2}{2\,\Delta^3}\big(w_{k+2} - 2\,w_{k+1} + 2\,w_{k-1} - w_{k-2}\big), \\ \frac{dw_k}{dt} - \frac{1}{2\,\Delta}\big(v_{k+1} - v_{k-1}\big) &= 0 \end{aligned} \tag{4.16}$$

for functions $w_k(t)$ and $v_k(t)$, where Δ denotes the spatial mesh-size. We integrate this system of ODEs using a fourth-order Runge–Kutta explicit scheme, taking as large a time-step τ as possible. We define $\lambda = \tau/\Delta$. Here we are interested in the continuous model (4.16) for small ϵ. (See [23] for results on numerical schemes.) The following figures may be taken to represent features of the continuous model (4.1): we carefully checked that reducing the mesh size further virtually does not change the numerical results.

The Riemann initial data for the numerical scheme is implemented as $(v_k(0), w_k(0)) = (v_l, w_l)$ for $k \le 0$ and (v_r, w_r) for $k > 0$. In Figures 4.1–4.2, we plot the numerical solution for several choices of initial data and parameters ϵ and α. From these figures,



(a)

(b)

Fig. 4.1. *Single nonclassical shock:* (a) *nonclassical* 1*-shock,* (b) *nonclassical* 2*-shock.*

we may compare the classical and nonclassical solutions. All the tests are performed on the interval $x \in [-3, 3]$ and with $m = 1$ in (3.2).

In Figure 4.1(a), we use the initial data $(v_l, w_l) = (1, 1)$ and $(v_r, w_r) = (-1.5, -2)$. The parameters are chosen to be $\Delta = 1/400$, $\lambda = .2$. The component w of the numerical solution is represented in Figure 4.1(a): the dashed line and the solid line correspond to $\alpha = 0$ and $\alpha = 10$, respectively. In the second case we do observe nonclassical behavior, i.e., a nonclassical 1-shock.

Figure 4.1(b) is similar to Figure 4.1(a), except that $(v_r, w_r) = (-1.25, 6)$. The dashed line represents a nonclassical 2-shock.

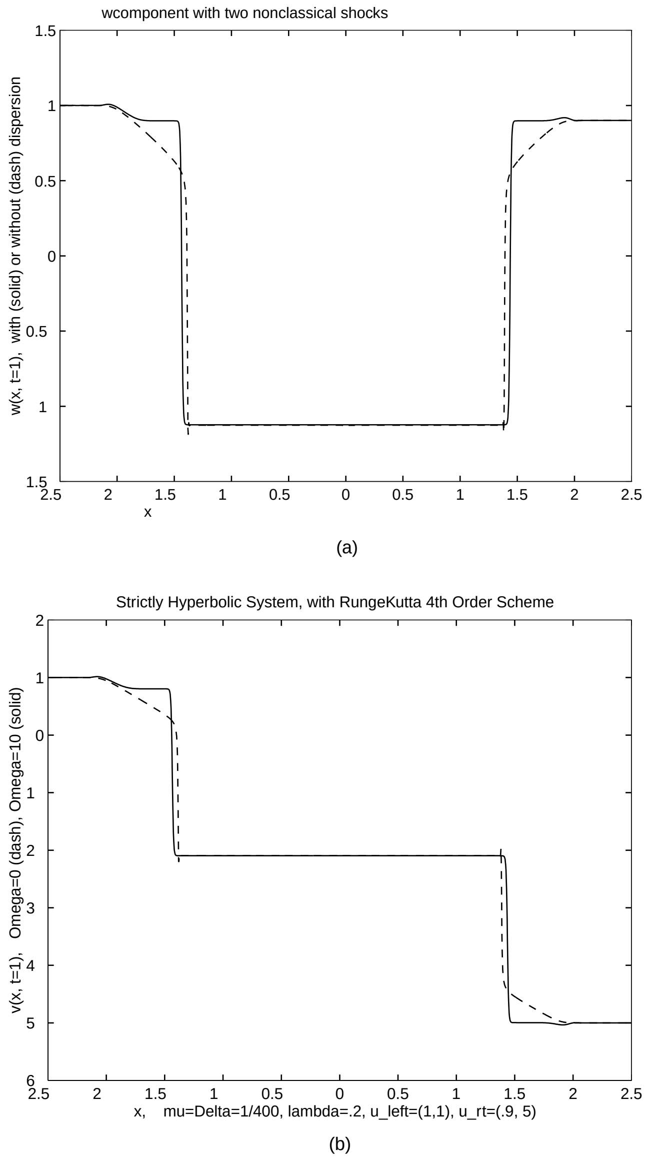
Figure 4.2 shows an example of a solution containing two nonclassical shocks, a 1-shock propagating in the left direction and a 2-shock going to the right. This is obtained with a suitable choice of the right state: $(v_r, w_r) = (.9, -5)$. The other parameters are the same as before. Figures 4.2(a) and 4.2(b) show the w- and v-component of the numerical solution, respectively.

5. Nonclassical shocks in magnetohydrodynamics.

5.1. Preliminaries. This section deals with a system, introduced by Freistühler [16], arising in the modeling of small amplitude solutions to conservation laws that are rotationally invariant:

$$\begin{aligned} &\partial_t v + \partial_x\big((v^2 + w^2)\, v\big) = 0, \\ &\partial_t w + \partial_x\big((v^2 + w^2)\, w\big) = 0. \end{aligned} \tag{5.1}$$

In magnetohydrodynamics, (v, w) represents transverse components of the magnetic field. This model is relevant to explain certain features observed in the solar wind around the Earth: Cohen and Kulsrud [8] and Wu and Kennel [65]. The model and its variants also arise in nonlinear elasticity. See also the interesting paper by Brio and Hunter [4]. The study of MHD traveling waves has a long history in the mathematical literature (consult, for instance, Conley and Smoller [9]). The system (5.1) has attracted attention of many researchers in recent years: Chen [6], Freistühler [17], Keyfitz and Kranzer [32], Liu and Wang [44], etc. Freistühler and Liu [19] established the nonlinear stability of overcompressive shocks for a parabolic regularization of the system (5.1).



(a)

(b)

FIG. 4.2. *Nonclassical shocks in both characteristic families:* (a) *w-component,* (b) *v-component.*

For smooth solutions, one can use polar coordinates

$$v = r\,\cos\theta, \qquad w = r\,\sin\theta, \qquad r \geq 0, \qquad \theta \in [0, 2\pi), \tag{5.2}$$

and rewrite (5.1) as

$$\partial_t r + \partial_x r^3 = 0, \tag{5.3}$$

$$\partial_t \theta + r^2\,\partial_x \theta = 0. \tag{5.4}$$

The equation (5.3) is a scalar conservation law with a nonconvex (cubic) flux. We deduce from (5.3) that $\lambda_2 = 3\,r^2$ is a wave speed for (5.1); it is the fast mode of the system and the corresponding characteristic field therefore fails to be genuine nonlinear. On the other hand, (5.4) is linearly degenerate since the slow mode wave speed $\lambda_1 = r^2$ is independent of θ.

Observe that the system is strictly hyperbolic everywhere but at the so-called *umbilic point* $v = w = 0$ or equivalently $r = 0$. The change of variable (5.2) is in fact ill defined at $r = 0$ since the angle θ may be arbitrary. The structure (5.3)–(5.4) reflects the property of invariance by rotation or isotropy of (5.1). There exists two main wave families:

- the *rotational discontinuities* keep the radius r constant while the angle θ may vary arbitrarily. Any entropy inequality would be satisfied by rotational discontinuities.
- the *fast shocks* keep the angle θ constant *modulo* π while the radius r may vary arbitrarily. An entropy inequality would select admissible fast shocks among all possible such shocks. Note that a rotational discontinuity always precedes a fast shock.

Consider now particular solutions to (5.1) such that $w = \rho\, v$, where ρ is a given constant. Such solutions will be called *coplanar* in this section. Then both equations in (5.1) reduce to the same equation,

$$\partial_t v + \left(1 + \rho^2\right) \partial_x v^3 \,=\, 0, \tag{5.5}$$

which is a scalar conservation law with cubic flux. Therefore, when the initial data for (5.1) are coplanar, one can attempt to solve the system (5.1) by solving the reduced equation (5.5). This is a saddle issue: the transformation $w = \rho\, v$ need not be compatible with a given regularization added to the right-hand side of (5.1). However, in several instances the solutions to (5.5) turn out to be relevant to describe the solutions to (5.1). Note finally that the "natural" entropy for (5.1),

$$U(v,w) = \frac{1}{2}\left(v^2 + w^2\right) = \frac{r^2}{2}, \qquad F(v,w) = \frac{3}{4}\left(v^2 + w^2\right)^2 = \frac{3\, r^4}{4}, \tag{5.6}$$

reduces, when $w = \rho\, v$, to an entropy pair for (5.5),

$$U(v) = \frac{1}{2} v^2, \qquad F(v,w) = \frac{3}{4}\left(1 + \rho^2\right) v^4, \tag{5.7}$$

which happens to be the one used in [22].

5.2. Unique admissible nonclassical entropy solution. The existence and properties of the nonclassical shocks for the cubic conservation law (5.3) were investigated in Hayes and LeFloch [22]. The equation (5.3), however, is supplemented with the constraint that $r \geq 0$, which prevents us from truly solving (5.3) independently of (5.4) for θ, even for coplanar initial data. The definitions in section 2 extend easily to (5.1) even though the system is not strictly hyperbolic. We are interested in solutions satisfying the single entropy inequality

$$\frac{1}{2}\partial_t\left(v^2 + w^2\right) + \frac{3}{4}\partial_x\left(v^2 + w^2\right)^2 \leq 0. \tag{5.8}$$

Our aim is to investigate the uniqueness of the nonclassical solutions for the system (5.1). Relying on the analysis in [22] we state, without proof, the following result.

THEOREM 5.1. *Consider the Riemann problem for the system* (5.1) *with initial data* (v_l, w_l) *and* (v_r, w_r). *When the data are noncoplanar, then there exists a unique* *solution to the Riemann problem satisfying the entropy inequality* (5.8): *it contains a rotational discontinuity connecting* (v_l, w_l) *to a point* (v_*, w_*) *with* $v_l^2 + w_l^2 = v_*^2 + w_*^2$ *followed by either a fast shock or a rarefaction connecting to* (v_r, w_r).

When the data are coplanar and the angles θ_l *and* θ_r *associated with the initial data satisfy* $\theta_r = \theta_r$ *(mod.* π*), the Riemann problem has a unique solution containing either a classical shock or a rarefaction.*

When the data are coplanar and $\theta_r = \pi + \theta_r$ *(mod.* π*), the Riemann problem admits a one-parameter family of entropy solutions containing a nonclassical shock connecting* (v_l, w_l) *to a point* (v_*, w_*) *with*

$$v_l^2 + w_l^2 \le v_*^2 + w_*^2 \le v_l^2 + w_l^2, \tag{5.9}$$

followed by either a fast shock or a rarefaction connecting to (v_r, w_r).

In the latter case, we can impose across the nonclassical shock a kinetic relation of the form

$$-s\,\frac{1}{2}\big[v^2 + w^2\big] + \frac{3}{4}\big[\big(v^2 + w^2\big)^2\big] = \varphi(s), \tag{5.10}$$

where the kinetic function $\varphi(s)$ *satisfies the property* $(s \ge 0)$

$$\begin{aligned} -\frac{3}{4}s^2 \;\le\; \varphi(s) \;\le\; 0, \\ \frac{d\varphi}{ds}(s) \;\le\; 0. \end{aligned} \tag{5.11}$$

A unique solution is selected by (5.10) *in the one-parameter family of solutions. This solution depends continuously (in the* L^1 *norm) on its end states for* coplanar *initial data.*

We refer to Hayes and LeFloch [22] for further details on the Riemann solution to the cubic conservation law (5.3). A solution to (5.1) using only classical shock waves always exists. We emphasize that the one-parameter family of solutions constructed in Theorem 5.1 includes as special cases the classical Riemann solution (defined from the Oleinik criterion) and the Riemann solution using a rotational discontinuity followed by a fast shock. For noncoplanar data, the Riemann solution constructed in Theorem 5.1 does not depend continuously upon its initial states. (Consider "quasi-coplanar" initial data.) It is conceivable that this lack of continuity may be related to physical instabilities in MHD fluid which cannot be fully described by the model (5.1).

The coplanar discontinuities connecting (r_L, θ_L) to (r, θ) with $r \in (0, r_L/2)$ and $\theta = \theta_L + \pi$ are *overcompressive* shock waves. They possess nonunique traveling wave profiles, due to the existence of a component θ. When viewed as shock to the underlying scalar cubic conservation law, they are *classical* shocks, however.

5.3. Convergence result. As we now demonstrate it numerically in section 5.4 below, the solutions found in Theorem 5.1 may arise from diffusive-dispersive regularizations of (5.1). We consider here the system $(\epsilon > 0,\ \alpha \in \mathbb{R})$

$$\begin{aligned} \partial_t v + \partial_x(v\,(v^2 + w^2)) = \epsilon\,\partial_{xx} v + \alpha\,\epsilon\,\partial_{xx} w, \\ \partial_t w + \partial_x(w\,(v^2 + w^2)) = \epsilon\,\partial_{xx} w - \alpha\,\epsilon\,\partial_{xx} v, \end{aligned} \tag{5.12}$$

called the derivative nonlinear Schrödinger–Burgers system. The right-hand side of (5.12) represents diffusive-dispersive effects arising in magnetic fluids due to the so-called Hall effect. When the ion inertia dispersion α can be neglected, (5.12) reduces to the Cohen–Kulsrud–Burgers (CKB) equations and converges, as $\epsilon \to 0$, to classical solutions. When $\alpha \neq 0$, the operator $\alpha\,\partial_{xx}$ in the right-hand side of (5.12) generates dispersion effect and nonclassical solutions may be obtained.

THEOREM 5.2. (1) *Let $\big(v^\epsilon, w^\epsilon\big)$ with $\alpha \in (-1,1)$ fixed be a family of solutions to* (5.12) *assuming at $t = 0$ a Cauchy data $\big(v_0^\epsilon, w_0^\epsilon\big)$ such that*

$$v_0^\epsilon, w_0^\epsilon \in L^2(\mathbb{R}) \cap L^4(\mathbb{R}) \tag{5.13}$$

uniformly in ϵ. Then $\big(v^\epsilon, w^\epsilon\big)$ is bounded in $L^\infty\big(\mathbb{R}_+, L^2(\mathbb{R}) \cap L^4(\mathbb{R})\big)$ and converges almost everywhere to a limiting function (v, w), a solution to (5.1) *in the sense of distributions.*

(2) *The pair $(U, F) = \big((v^2 + w^2)/2,\, 3\,(v^2 + w^2)^2/4\big)$ is compatible in the sense* (2.3) *with the diffusive-dispersive regularization* (5.12). *Limits of traveling wave solutions to* (5.12) *additionally satisfy the entropy inequality* (5.8).

Proof of Theorem 5.2. The proof relies on the compensated compactness method of DiPerna [13] and more specifically the results in Chen [6]. We restrict attention to deriving the main a priori estimates needed in applying the theory, referring to [6, 13] for the details. The following entropy balance follows by multiplying the equations in (5.12) by v and w, respectively:

$$\begin{aligned}\tfrac{1}{2}\partial_t\big(v^2 + w^2\big) + \tfrac{3}{4}\partial_x\big(v^2 + w^2\big)^2 \;=\; & -\epsilon\,|\partial_x v|^2 - \epsilon\,|\partial_x w|^2 \\ & + \epsilon\,\partial_x\big(v\,\partial_x v + w\,\partial_x w\big) + \alpha\,\epsilon\,\partial_x\big(v\,\partial_x w - w\,\partial_x v\big).\end{aligned} \tag{5.14}$$

Integrating (5.14) over $(0, T) \times \mathbb{R}$ yields

$$\int_{\mathbb{R}} \frac{1}{2}\big(v^2 + w^2\big)(T)\,dx + \int_0^T \int_{\mathbb{R}} \epsilon\,\big(|\partial_x v|^2 + |\partial_x w|^2\big)\,dxdt \;\le\; \int_{\mathbb{R}} \frac{1}{2}\big(v^2 + w^2\big)(0)\,dx \;\le\; O(1). \tag{5.15}$$

Observe that the dispersive terms canceled out in this derivation, so that the estimate (5.15) does not depend on the coefficient $\alpha \in \mathbb{R}$.

We now multiply (5.14) on both sides by $v^2 + w^2$ and integrate over $\mathbb{R}$ to get

$$\begin{aligned}&\frac{d}{dt}\int_{\mathbb{R}} \frac{1}{4}\big(v^2 + w^2\big)^2\,dx + \int_{\mathbb{R}} \frac{1}{2}\partial_x\big(v^2 + w^2\big)^3\,dx \\ &\quad = -\int_{\mathbb{R}} \epsilon\,\big(v^2 + w^2\big)\,\big(|\partial_x v|^2 + |\partial_x w|^2\big)\,dx - \int_{\mathbb{R}} \epsilon\,\big|\partial_x\big(v^2 + w^2\big)\big|^2\,dx \\ &\qquad + \int_{\mathbb{R}} \alpha\,\epsilon\,\big(v^2 + w^2\big)\,\partial_x\big(v\,\partial_x w - w\,\partial_x v\big)\,dx.\end{aligned} \tag{5.16}$$

Thus we obtain

$$\begin{aligned}&\int_{\mathbb{R}} \frac{1}{4}\big(v^2 + w^2\big)^2(T)\,dx + \int_0^T \int_{\mathbb{R}} \epsilon\,\big(v^2 + w^2\big)\,\big(|\partial_x v|^2 + |\partial_x w|^2\big)\,dxdt \\ &\quad + \int_0^T \int_{\mathbb{R}} \epsilon\,\big|\partial_x\big(v^2 + w^2\big)\big|^2\,dxdt \\ &\quad = \int_{\mathbb{R}} \frac{1}{4}\big(v^2 + w^2\big)^2(0)\,dx + \int_0^T \int_{\mathbb{R}} \alpha\,\epsilon\,\big(v^2 + w^2\big)\,\big(v\,\partial_x w - w\,\partial_x v\big)\,dxdt.\end{aligned} \tag{5.17}$$

When $|\alpha| < 1$, the integrand of the last term in the right-hand side of (5.17) can be estimated by integrands of the left-hand side, namely,

$$\begin{aligned}\left|\alpha\,\epsilon\,\partial_x\big(v^2+w^2\big)\,\big(v\,\partial_x w - w\,\partial_x v\big)\right| &\le \frac{\alpha}{2}\,\epsilon\,\left|\partial_x\big(v^2+w^2\big)\right|^2 + \frac{\alpha}{2}\,\epsilon\,\left|v\,\partial_x w - w\,\partial_x v\right|^2\\ &\le \frac{\alpha}{2}\,\epsilon\,\left|\partial_x\big(v^2+w^2\big)\right|^2 + \alpha\,\epsilon\,\left|v\,\partial_x w\right|^2 + \alpha\,\epsilon\,\left|w\,\partial_x v\right|^2\\ &\le \frac{\alpha}{2}\,\epsilon\,\left|\partial_x\big(v^2+w^2\big)\right|^2\\ &\qquad + \alpha\,\epsilon\,\big(v^2+w^2\big)\,\big(|\partial_x v|^2 + |\partial_x w|^2\big).\end{aligned}$$

Therefore (5.17) implies

$$\begin{aligned}\int_{\mathbb{R}} \frac{1}{4}\big(v^2+w^2\big)^2(T)\,dx + \int_0^T \int_{\mathbb{R}} \big(1-\alpha/2\big)\,\epsilon\,\big(v^2+w^2\big)\,\big(|\partial_x v|^2 + |\partial_x w|^2\big)\,dxdt\\ + \int_0^T \int_{\mathbb{R}} \big(1-\alpha\big)\,\epsilon\,\left|\partial_x\big(v^2+w^2\big)\right|^2\,dxdt \;\le\; 0.\end{aligned} \tag{5.18}$$

The estimates (5.15) and (5.18) provide L^p uniform bounds for v^ϵ and w^ϵ, together with some derivative estimates. These estimates can be used along the lines of the proof in Schonbek [53] (and [22]) to show that a Young measure associated with (v^ϵ, w^ϵ) satisfies Tartar's commutation equation. The reduction theorem in [6] may be extended to L^p Young measures and shows that

$$\begin{aligned}&v^\epsilon \to v, \qquad w^\epsilon \to w &&\text{in the weak sense,}\\ &v_\epsilon^2 + w_\epsilon^2 \to v^2 + w^2 &&\text{in the strong sense.}\end{aligned} \tag{5.19}$$

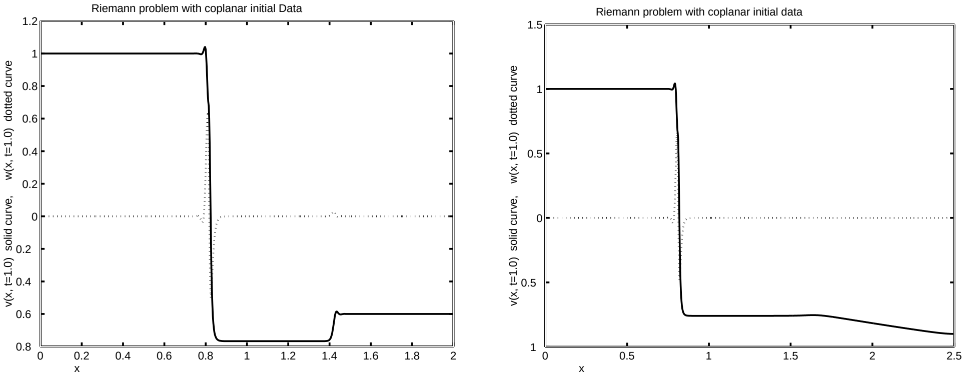
One can pass to the limit in (5.12) and deduce (5.1) as $\epsilon \to 0$.

Item (2) of the theorem follows from (5.14) and the uniform estimates (5.18). Observe that the first two terms in the right-hand side of (5.14) are nonpositive and converge to nonnegative bounded measures. The third term converges to zero in the sense of distributions. On the other hand the last term in the right-hand side of (5.14), due to the dispersive terms in (5.12), does not contribute to the dissipation measure (for the quadratic entropy only); namely, for each smooth function θ with compact support, one has

$$\begin{aligned}\left|\int_0^T \int_{\mathbb{R}} \epsilon\,\partial_x\big(v\,\partial_x w - w\,\partial_x v\big)\,\theta\,dxdt\right| &\le \int_0^T \int_{\mathbb{R}} \epsilon\,\left|v\,\partial_x w\right|\,\partial_x\theta\,dxdt\\ &\quad + \int_0^T \int_{\mathbb{R}} \epsilon\,\left|w\,\partial_x v\right|\,\partial_x\theta\,dxdt\\ &\le O(1)\,\epsilon\,\|v\,\partial_x w\|_{L^2\big((0,T)\times\mathbb{R}\big)}\\ &\quad + O(1)\,\epsilon\,\|v\,\partial_x w\|_{L^2\big((0,T)\times\mathbb{R}\big)}\\ &\le O(1)\,\epsilon^{1/2} \;\to\; 0.\end{aligned}$$

This completes the proof of Theorem 5.2. □

5.4. Numerical experiments. For coplanar initial data, we numerically demonstrate the existence of nonclassical shocks. We employ the following semidiscrete



(a) (b)

FIG. 5.1. *The slow shock is nonclassical: it cannot be a rotational wave, since across this shock,* $|u_m|^2 < |u_l|^2$.

approximation to the system (5.12):

$$
\begin{aligned}
&\frac{dv_k}{dt} + \frac{1}{2\,\Delta}\big(v_{k+1}\,(v_{k+1}^2 + w_{k+1}^2) - v_{k-1}\,(v_{k-1}^2 + w_{k-1}^2)\big)\\
&\qquad\qquad = \frac{\epsilon}{\Delta^2}\big(v_{k+1} - 2\,v_k + v_{k-1}\big) + \frac{\alpha\,\epsilon}{\Delta^2}\big(w_{k+1} - 2\,w_k + w_{k-1}\big),\\
&\frac{dw_k}{dt} + \frac{1}{2\,\Delta}\big(w_{k+1}\,(v_{k+1}^2 + w_{k+1}^2) - w_{k-1}\,(v_{k-1}^2 + w_{k-1}^2)\big)\\
&\qquad\qquad = \frac{\epsilon}{\Delta^2}\big(w_{k+1} - 2\,w_k + w_{k-1}\big) - \frac{\alpha\,\epsilon}{\Delta^2}\big(v_{k+1} - 2\,v_k + v_{k-1}\big)
\end{aligned}
\tag{5.20}
$$

for functions $v_k(t)$ and $w_k(t)$, where Δ denotes the spatial mesh-size. We integrate this system of ODEs in the same fashion as in subsection 4.2. The Riemann initial data for the numerical scheme are implemented as $(v_k(0), w_k(0)) = (v_l, w_l)$ for $k \le 0$ and (v_r, w_r) for $k > 0$.

In Figure 5.1, we plot the numerical results for two different coplanar data. The parameters are chosen to be $\Delta = 1/400$, $\epsilon = 1/800$, and $\alpha = 5/2$. In Figure 5.1(a), we use the initial data $(v_l, w_l) = (1, 0)$ and $(v_r, w_r) = (-.6, 0)$. The solid and the dotted lines represent the v- and w-components of the solution at the time $t = 1$, respectively. In Figure 5.1(b), we picture the results obtained with, instead, $(v_r, w_r) = (-.85, 0)$.

Appendix: Proof of Lemma 2.3. We follow Liu in [42] and treat the case (2.10a). The case (2.10b) is entirely similar. The statement on the wave speed follows easily from our assumption that $\nabla\lambda_j \cdot r_j$ changes sign only once along a shock curve. Let us show that the shock speed satisfies similar properties. By differentiating the Rankine–Hugoniot relation (2.13), we get

$$
-\frac{\partial}{\partial\mu_j}\bar\lambda_j(u_0, w_j)(w_j - u_0) + \big(Df(w_j) - \bar\lambda_j(u_0, w_j)\big)\frac{dw_j}{d\mu_j} \;=\; 0. \tag{A.1}
$$

Using the decompositions

$$
w_j - u_0 = \sum_{k=1}^{N} \alpha_k(u_0, w_j)\, r_k(w_j)
$$

and

$$\frac{dw_j}{d\mu_j} = \sum_{k=1}^{N} \beta_k(u_0, w_j)\, r_k(w_j),$$

we deduce that, for $k = 1, 2, \dots, N$,

$$-\frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j)\, \alpha_k(u_0, w_j) + \big(\lambda_k(w_j) - \bar{\lambda}_j(u_0, w_j)\big)\, \beta_k(u_0, w_j) \;=\; 0.$$

In particular, for $k = j$,

$$\frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j)\, \alpha_j(u_0, w_j) = \big(\lambda_j(w_j) - \bar{\lambda}_j(u_0, w_j)\big)\, \beta_j(u_0, w_j).$$

In view of our assumption (2.21), the coefficient $\alpha_j(u_0, w_j) = l_j(w_j) \cdot (w_j - u_0)$ has the same sign as $\mu_j - \mu_j(u_0)$, while $\beta_j(u_0, w_j) = l_j(w_j) \cdot dw_j / d\mu_j$ is strictly positive. Therefore for $\mu_j > \mu_j(u_0)$ we have

$$\begin{aligned} &\frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j) = 0 \text{ (resp., } > 0 \text{ or } < 0\big) \quad \text{iff} \\ &\lambda_j(w_j) - \bar{\lambda}_j(u_0, w_j) = 0 \text{ (resp., } < 0 \text{ or } > 0\big), \end{aligned} \tag{A.2}$$

while for $\mu_j < \mu_j(u_0)$ the reversed inequalities are satisfied. Moreover it follows from (A.1) that (up to a multiplicative factor)

$$\frac{dw_j}{d\mu_j} \;=\; r_j(w_j) \quad \text{if} \quad \frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j) \;=\; 0. \tag{A.3}$$

Denote by $\mu_j^\star(u_0)$ a point achieving the equality in (A.2). We now prove that, at the critical point $\mu_j = \mu_j^\star(u_0)$,

$$\begin{aligned} &\frac{\partial^2}{\partial \mu_j^2} \bar{\lambda}_j(u_0, w_j) = 0 \text{ (resp., } > 0 \text{ or } < 0\big) \quad \text{iff} \\ &\nabla \lambda_j(w_j) \cdot r_j(w_j) = 0 \,\text{(resp., } < 0 \text{ or } > 0\big) \end{aligned} \tag{A.4}$$

if $\mu_j^\star(u_0) > \mu_j(u_0)$, while the reversed inequalities are satisfied if $\mu_j^\star(u_0) < \mu_j(u_0)$. Namely, first rewrite the relation (A.1) (by using (A.3)) in the form

$$\begin{aligned} &\big(Df(w_j) - \bar{\lambda}_j(u_0, w_j)\big) \left(\frac{dw_j}{d\mu_j} - r_j(w_j)\right) \\ &\quad = \frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j)(w_j - u_0) - \big(\lambda_j(w_j) - \bar{\lambda}_j(u_0, w_j)\big)\, r_j(w_j). \end{aligned} \tag{A.5}$$

Differentiating (A.5) once more, we obtain

$$\begin{aligned} &\frac{\partial}{\partial \mu_j} \big(Df(w_j) - \bar{\lambda}_j(u_0, w_j)\big) \left(\frac{dw_j}{d\mu_j} - r_j(w_j)\right) \\ &+ \big(Df(w_j) - \bar{\lambda}_j(u_0, w_j)\big) \left(\frac{d^2 w_j}{d\mu_j^2} - \frac{\partial}{\partial \mu_j} r_j(w_j)\right) \\ &= \frac{\partial^2}{\partial \mu_j^2} \bar{\lambda}_j(u_0, w_j)(w_j - u_0) + \frac{\partial}{\partial \mu_j} \bar{\lambda}_j(u_0, w_j) \frac{dw_j}{d\mu_j} \\ &- \frac{\partial}{\partial \mu_j} \big(\lambda_j(w_j) - \bar{\lambda}_j(u_0, w_j)\big)\, r_j(w_j) - \big(\lambda_j(w_j) - \bar{\lambda}_j(u_0, w_j)\big) \frac{\partial}{\partial \mu_j} r_j(w_j). \end{aligned}$$

Plugging the value $\mu_j = \mu_j^\star(u_0)$ in the above formula and using (A.2)–(A.3), we obtain

$$\begin{aligned}\big(Df(w_j) - \bar\lambda_j(u_0, w_j)\big)\left(\frac{d^2 w_j}{d\mu_j^2} - \frac{\partial}{\partial\mu_j} r_j(w_j)\right) &= \frac{\partial^2}{\partial\mu_j^2}\bar\lambda_j(u_0, w_j)(w_j - u_0)\\ &\quad - \frac{\partial}{\partial\mu_j}\lambda_j(w_j)\, r_j(w_j).\end{aligned}$$

Multiplying the latter by $l_j(w_j)$ and observing that $\bar\lambda_j(u_0, w_j) = \lambda_j(w_j)$ so that the left-hand side vanishes, we arrive at

$$\begin{aligned}\frac{\partial^2}{\partial\mu_j^2}\bar\lambda_j(u_0, w_j)\, l_j(w_j)\cdot(w_j - u_0) &= \frac{\partial}{\partial\mu_j}\lambda_j(w_j)\\ &= \nabla\lambda_j(w_j)\cdot r_j(w_j).\end{aligned}$$

The desired result (A.4) follows immediately from the above formula and assumption (2.21ii).

We now use the notation $g(\mu_j) := \lambda_j(w_j(\mu_j; u_0))$ and $h(\mu_j) := \bar\lambda_j(u_0, w_j(\mu_j; u_0))$. The property (A.2) shows that (2.24a) is satisfied for values μ_j close enough to $\mu_j(u_0)$, at least. Consider the largest value $\mu_j < \mu_j(u_0)$ such that $h(\zeta_j) - g(\zeta_j) > 0$ holds for all $\zeta \in (\mu_j, \mu_j(u_0)$. Call this value $\mu_j^\star(u_0)$ and observe that $h(\mu_j^\star(u_0)) = g(\mu_j^\star(u_0))$. In view of (A.2) one also has $h'(\mu_j^\star(u_0)) = 0$.

Assume that $\mu_j^\star(u_0) > 0$. In view of (A.4), one has $h''(\mu_j^\star(u_0)) > 0$ since $\mu_j^\star(u_0) > 0$. Thus the function should decrease for $\mu_j < \mu_j^\star(u_0)$ at least in a small neighborhood of $\mu_j^\star(u_0)$. According to (A.2), the wave speed should then be above the shock speed in this range, and so the wave speed g should be nonincreasing. The function g is increasing near $\mu_j(u_0)$ and nonincreasing near $\mu_j^\star(u_0)$, so g must have a critical point in the interval $\big[\mu_j^\star(u_0), \mu_j(u_0)\big)$. Since the only critical point of the wave speed is $\mu_j = 0$ and $\mu_j^\star(u_0) > 0$ by assumption, we reach a contradiction. Henceforth, one must have $\mu_j^\star(u_0) \le 0$.

Finally the shock speed is monotone in the whole region $\mu_j < \mu_j^\star(u_0)$, since otherwise that would imply the existence of a critical point for the function g, which is not possible. This completes the proof of Lemma 2.3. □

Acknowledgments. Part of this work was carried out while the second author was visiting the Institute for Mathematics and Its Applications (IMA), Minneapolis, MN, in September 1995 and January 1996.

The authors learned that D. Marchesin, B. Plohr, and K. Zumbrun recently observed that undercompressive shocks may arise in strictly hyperbolic systems. In a recent and interesting preprint by H. Freistühler, the selection of undercompressive shocks in hyperbolic systems and their stability in the multidimensional setting is also investigated.

REFERENCES

[1] R. Abeyaratne and J.K. Knowles, *Kinetic relations and the propagation of phase boundaries in solids*, Arch. Rational Mech. Anal., 114 (1991), pp. 119–154.

[2] R. Abeyaratne and J.K. Knowles, *Implications of viscosity and strain-gradient effects for the kinetics of propagating phase boundaries in solids*, SIAM J. Appl. Math., 51 (1991), pp. 1205–1221.

[3] A. Azevedo, D. Marchesin, B.J. Plohr, and K. Zumbrun, *Non-uniqueness of solutions of Riemann problems caused by* 2*-cycles of shock waves*, in Proceedings of the Fifth International Conference on Hyperbolic Problems: Theory, Numerics, Applications, J. Glimm,

M.J. Graham, J.W. Grove, and B.J. Plohr, eds., World Scientific Editions, River Edge, N.J., 1996, pp. 43–51.

[4] M. Brio and J. Hunter, *Rotationally invariant hyperbolic waves*, Comm. Pure Appl. Math., 43 (1990), pp. 1037–1053.

[5] M. Brio and C.C. Wu, *An upwind differencing scheme for the equations of ideal magnetohydrodynamics*, J. Comput. Phys., 75 (1988), pp. 400–422.

[6] G.Q. Chen, *Hyperbolic systems of conservation laws with a symmetry*, Comm. Partial Differential Equations, 16 (1991), pp. 1461–1467.

[7] 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.

[8] R.H. Cohen and R.M. Kulsrud, *Non-linear evolution of quasi-parallel propagating hydromagnetic waves*, Phys. Fluid, 17 (1974), pp. 2215–2225.

[9] C. Conley and J.A. Smoller, *On the structure of the magnetohydrodynamics shock waves*, Comm. Pure Appl. Math., 28 (1974), pp. 367–375.

[10] C.M. Dafermos, *The entropy rate admissibility criterion for solutions of hyperbolic conservation laws*, J. Differential Equations, 14 (1973), pp. 202–212.

[11] C.M. Dafermos, *Hyperbolic systems of conservation laws*, in Proceedings 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.

[12] C.M. Dafermos, *Admissible fans in nonlinear hyperbolic systems*, Arch. Rational Mech. Anal., 106 (1989), pp. 243–260.

[13] R.J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27–70.

[14] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, Ann Arbor, London, 1992.

[15] H.T. Fan and M. Slemrod, *The Riemann problem for systems of conservation laws of mixed type*, in Shock Induced Transitions and Phase Structures in General Media, R. Fosdick, E. Dunn, and H. Slemrod, eds., IMA Vol. Math. Appl. 52, Springer-Verlag, 1993, pp. 61–91.

[16] H. Freistühler, *Anomale Schocks, strukturell labile Lösungen und die Geometrie der Rankine-Hugoniot-Bedingungen*, Ph.D. thesis, Ruhr-Univ. Bochum, Germany, 1987.

[17] H. Freistühler, *Dynamical stability and vanishing viscosity: A case study of a non-strictly hyperbolic system of conservation laws*, Comm. Pure Appl. Math., 45 (1992), pp. 561–582.

[18] H. Freistühler, *Stability of nonclassical shock waves*, in Proceedings of the Fifth International Conference on Hyperbolic Problems: Theory, Numerics, Applications, J. Glimm, M.J. Graham, J.W. Grove, and B.J. Plohr, eds., World Scientific Editions, River Edge, N.J., 1996, pp. 120–129.

[19] H. Freistühler and T.P. Liu, *Nonlinear stability of overcompressive shock waves in a rotationally invariant system of viscous conservation laws*, Comm. Math. Phys., 153 (1993), pp. 147–158.

[20] J. Glimm, *Nonlinear waves: Overview and problems*, in Multidimensional Hyperbolic Problems and Computations, J. Glimm and A. Majda, eds., IMA Vol. Math. Appl. 29, Springer-Verlag, New York, 1991.

[21] B.T. Hayes and P.G. LeFloch, *Measure-solutions to a strictly hyperbolic system of conservation laws*, Nonlinearity, 9 (1996), pp. 1547–1563.

[22] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Scalar conservation laws*, Arch. Rational Mech. Anal., 139 (1997), pp. 1–56.

[23] B.T. Hayes and P.G. LeFloch, *Nonclassical shocks and kinetic relations: Finite difference schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 2169–2194.

[24] B.T. Hayes and P.G. LeFloch, in preparation.

[25] T.Y. Hou, P. Rosakis, and P.G. LeFloch, *A level-set approach to the computation of twinning and phase transition dynamics*, J. Comput. Phys., 150 (1999), pp. 302–331.

[26] L. Hsiao, *Uniqueness of admissible solutions of Riemann problems of systems of conservation laws of mixed type*, J. Differential Equations, 86 (1990), pp. 197–233.

[27] E. Isaacson, D. Marchesin, C.F. Palmeira, and B.J. Plohr, *A global formalism for nonlinear waves in conservation laws*, Comm. Math. Phys., 146 (1992), pp. 505–552.

[28] E. L. Isaacson, D. Marchesin, and B. J. Plohr, *Transitional waves for conservation laws*, SIAM J. Math. Anal., 21 (1990), pp. 837–866.

[29] E. Isaacson and B. Temple, *Nonlinear resonance in systems of conservation laws*, SIAM J. Appl. Math., 52 (1992), pp. 1260–1278.

[30] D. Jacobs, W.R. McKinney, and M. Shearer, *Traveling wave solutions of the modified Korteweg-deVries Burgers equation*, J. Differential Equations, 116 (1995), pp. 448–467.

[31] B. Keyfitz, *A geometric theory of conservation laws which change type*, Z. Angew. Math.

Mech., 75 (1995), pp. 571–581.

[32] B. Keyfitz and H. Kranzer, *A system of nonstrictly hyperbolic conservation arising in elasticity theory*, Arch. Rational Mech. Anal., 72 (1980), pp. 219–241.

[33] A. Kulikovsky, Dokl. Acad. Nauk. USSR, 275 (1984), pp. 1349–1352 (in Russian).

[34] P.D. Lax, *Hyperbolic systems of conservation laws*, II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[35] P.D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM, Philadelphia, PA, 1973.

[36] P.D. Lax, *The zero dispersion limit, a deterministic analogue of turbulence*, Comm. Pure Appl. Math., 44 (1991), pp. 1047–1056.

[37] P.D. Lax and C.D. Levermore, *The small dispersion limit of the Korteweg-deVries equation*, Comm. Pure Appl. Math., 36 (1983) I, pp. 253–290, II, pp. 571–593, III, pp. 809–829.

[38] P.G. LeFloch, *Propagating phase boundaries: Formulation of the problem and existence via the Glimm scheme*, Arch. Rational Mech. Anal., 123 (1993), pp. 153–197.

[39] P.G. LeFloch and R. Natalini, *Conservation laws with vanishing nonlinear diffusion and dispersion*, Nonlinear Anal., 36 (1999), pp. 213–230.

[40] P.G. LeFloch and A.E. Tzavaras, *Nonconservative hyperbolic systems: Existence theory for the Riemann problem*, in preparation.

[41] T.P. Liu, *The Riemann problem for general* 2×2 *conservation laws*, Trans. Amer. Math. Soc., 199 (1974), pp. 89–112.

[42] T.P. Liu, *Admissible solutions of hyperbolic conservation laws*, Mem. Amer. Math. Soc. 30, 1981.

[43] T.P. Liu, *Nonlinear stability and instability of overcompressive shock waves*, IMA Vol. Math. Appl. 52, Springer-Verlag, New York, 1993, pp. 159–167.

[44] T.P. Liu and J.H. Wang, *On a nonstrictly hyperbolic system of conservation laws*, J. Differential Equations, 57 (1985), pp. 1–14.

[45] T.P. Liu and Z. Xin, *Stability of viscous shock waves associated with a nonstrictly hyperbolic system*, Comm. Pure Appl. Math., 45 (1992), pp. 361–388.

[46] T.P. Liu and K. Zumbrun, *Nonlinear stability of an undercompressive shock for complex Burgers equation*, Comm. Math. Phys., 168 (1995), pp. 163–186.

[47] T.P. Liu and K. Zumbrun, *On nonlinear stability of general undercompressive viscous shock waves*, Comm. Math. Phys., 174 (1995), pp. 319–345.

[48] R. Menikoff and B.J. Plohr, *The Riemann problem for fluid flow of real materials*, Rev. Modern Phys., 61 (1989), pp. 75–130.

[49] F. Murat, *A survey on compensated compactness*, in Contributions to Modern Calculus of Variation, L. Cesari, ed., Pitman Res. Notes Math. Ser. 148, Longman, Harlow, UK, 1987, pp. 145–183.

[50] O. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Uspekhio Mat. Nauk. (N.S.), 12 (1957), pp. 3–73; Amer. Math. Transl. Ser. 2, 26, pp. 95–172 (in English).

[51] S. Schecter, D. Marchesin, and B.J. Plohr, *Structurally stable Riemann solutions*, J. Differential Equations, 126 (1996), pp. 303–354.

[52] S. Schecter and M. Shearer, *Undercompressive shocks for nonstrictly hyperbolic conservation laws*, Dynamics Differential Equations, 3 (1991), pp. 199–271.

[53] M.E. Schonbek, *Convergence of solutions to nonlinear dispersive equations*, Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.

[54] D. Serre and J.W. Shearer, *Convergence with physical viscosity for nonlinear elasticity*, unpublished notes.

[55] J.W. Shearer, *Global existence and compactness in* L^p *for the quasilinear wave equation*, Ph.D. thesis, University of California, Berkeley, CA, 1991.

[56] M. Shearer, *The Riemann problem for a class of conservation laws of mixed type*, J. Differential Equations, 46 (1982), pp. 426–443.

[57] M. Shearer and Y. Yang, *The Riemann problem for the p-system of conservation laws of mixed type with a cubic nonlinearity*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 675–699.

[58] M. Shearer, D.G. Schaeffer, D. Marchesin, and P. Paes-Leme, *Solution of the Riemann problem for a prototype* 2×2 *system of nonstrictly hyperbolic conservation laws*, Arch. Rational Mech. Anal., 97 (1987), pp. 299–320.

[59] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[60] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J.M. Ball, ed., NATO Adv. Sci. Inst. Ser. C. Math. Phys., 1983, pp. 263–285.

[61] L. Truskinovsky, *Dynamics of non-equilibrium phase boundaries in a heat conducting nonlinear elastic medium*, J. Appl. Math. Mech., 51 (1987), pp. 777–784.

[62] L. Truskinovsky, *Kinks versus shocks*, in Shock Induced Transitions and Phase Structures in General Media, R. Fosdick, E. Dunn, and H. Slemrod, eds., IMA Vol. Math. Appl. 52, Springer-Verlag, New York, 1993, pp. 185–229.

[63] C.C. Wu, *Formation, structure, and stability of MHD intermediate shocks*, J. Geophys. Res., 95 (1990), pp. 8149–8175.

[64] C.C. Wu, *New theory of MHD shock waves*, in Viscous Profiles and Numerical Methods for Shock Waves, M. Shearer, ed., SIAM, Philadelphia, PA, 1991, pp. 209–236.

[65] C.C. Wu and C.F. Kennel, *The small amplitude magnetohydrodynamic Riemann problem*, Phys. Fluids B, 5 (1993), pp. 2877–2886.

SIAM J. MATH. ANAL.
Vol. 31, No. 5, pp. 992–1029

UNIFORM ASYMPTOTIC FORMULA FOR ORTHOGONAL POLYNOMIALS WITH EXPONENTIAL WEIGHT*

W.-Y. QIU† AND R. WONG‡

Abstract. Let $\{p_n(x)\}_{n\geq 0}$ be the set of orthonormal polynomials with respect to the exponential weight $w(x) = e^{-v(x)}$, where $v(x) = x^{2m} + \cdots$ is a monic polynomial of degree $2m$ with $m \geq 2$ and is even. An asymptotic approximation is obtained for $p_n(x)$, as $n \to \infty$, which holds uniformly for $0 \leq x \leq O(n^{1/2m})$. As a corollary, a three-term asymptotic expansion is also derived for the zeros of these polynomials.

Key words. orthogonal polynomials, exponential weight, uniform asymptotic approximation, turning point, zeros

AMS subject classifications. 41A60, 33C45

PII. S0036141098344671

1. Introduction.

Let $\{p_n(x)\}_{n\geq 0}$ be the set of orthonormal polynomials with respect to the exponential weight function

$$w(x) = e^{-v(x)}, \qquad x \in (-\infty, \infty), \tag{1.1}$$

where $v(x) = x^{2m} + \cdots$ is a monic polynomial of degree $2m$ with $m \geq 2$. That is, the polynomials $p_n(x)$ satisfy

$$\int_{-\infty}^{\infty} p_n(x)p_k(x)w(x)dx = \begin{cases} 1, & n = k, \\ 0, & n \neq k. \end{cases} \tag{1.2}$$

Let $\gamma_n > 0$ denote the leading coefficient of $p_n(x), n = 0, 1, 2, \ldots$. Then the polynomials satisfy the three-term recurrence relation

$$(x - b_n)p_n(x) = a_{n+1}p_{n+1}(x) + a_n p_{n-1}(x), \tag{1.3}$$

where $a_0 = 0, a_n = \gamma_{n-1}/\gamma_n$ and

$$b_n = \int_{-\infty}^{\infty} x p_n^2(x)w(x)dx, \qquad n = 0, 1, 2, \ldots. \tag{1.4}$$

In [6], Nevai studied the asymptotic behavior of $p_n(x)$ when $w(x) = e^{-x^4}$ and obtained a Plancherel–Rotach type asymptotic formula for these polynomials. The behavior of their zeros has been investigated by Máté, Nevai, and Totik [4]. More precisely, they have derived a two-term expansion for the largest zero of $p_n(x)$. Similar results have been obtained by Sheen [8], when $w(x) = e^{-x^6}$. Recently, Bo and Wong [2] reconsidered the orthogonal polynomial $p_n(x)$ when $w(x) = e^{-x^4}$. They constructed an asymptotic formula which holds uniformly in an interval containing even the critical

*Received by the editors September 18, 1998; accepted for publication (in revised form) June 23, 1999; published electronically April 20, 2000.
http://www.siam.org/journals/sima/31-5/34467.html

†Department of Mathematics, Fudan University, Shanghai, People's Republic of China.

‡Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (mawong@cityu.edu.hk). The research of this author was partially supported by grants from the Research Grant Council of Hong Kong.

value $x = (4n/3)^{1/4}$. As an application of their result, they derived a four-term asymptotic expansion for the kth zero of $p_n(x)$ for any fixed $k = 1, 2, \ldots, n$.

For the weight function (1.1) with $v(x)$ being convex, Chen and Ismail [3] proved that $p_n(x)$ satisfies the differential recurrence relation

$$p_n'(x) = -B_n(x)p_n(x) + A_n(x)p_{n-1}(x), \qquad n = 0, 1, 2, \ldots, \tag{1.5}$$

where

$$A_n(x) = a_n \int_{-\infty}^{\infty} \frac{v'(x) - v'(y)}{x - y} p_n^2(y)w(y)dy \tag{1.6}$$

and

$$B_n(x) = a_n \int_{-\infty}^{\infty} \frac{v'(x) - v'(y)}{x - y} p_n(y)p_{n-1}(y)w(y)dy. \tag{1.7}$$

Note that if $v(x)$ is even, then it has the form

$$v(x) = \sum_{k=0}^{m} v_k x^{2k}, \qquad v_m = 1, \qquad m \geq 2, \tag{1.8}$$

and we have

$$\frac{v'(x) - v'(y)}{x - y} = 2\sum_{k=1}^{m} k v_k \sum_{l=0}^{2k-2} x^{2k-2-l} y^l. \tag{1.9}$$

In this paper, we shall assume that $v(x)$ is a convex and even polynomial of the form (1.8), which is also the main case considered by Chen and Ismail [3, Theorem 4.3 and section 5]. It is easily seen that when $v(x)$ is convex, the function $A_n(x)$ in (1.6) is positive. Furthermore, when $v(x)$ is of the form (1.8), we have $b_n = 0$ by (1.4), and the recurrence relation (1.3) becomes

$$xp_n(x) = a_{n+1}p_{n+1}(x) + a_n p_{n-1}(x). \tag{1.10}$$

One of the main result in [3] is that the function

$$Y_n(x) = \frac{p_n(x)}{\sqrt{A_n(x)}}\, e^{-v(x)/2} \tag{1.11}$$

satisfies the second-order differential equation

$$Y''(x) + V(x, n)Y(x) = 0, \tag{1.12}$$

where

$$\begin{aligned} V(x,n) = {} & A_n(x)A_{n-1}(x)\frac{a_n}{a_{n-1}} + B_n'(x) + \frac{1}{2}v''(x) + \frac{1}{2}\frac{A_n''(x)}{A_n(x)} \\ & - \left[B_n(x) + \frac{1}{2}v'(x)\right]^2 - \frac{A_n'(x)}{A_n(x)}\left[B_n(x) + \frac{1}{2}v'(x)\right] - \frac{3}{4}\left[\frac{A_n'(x)}{A_n(x)}\right]^2. \end{aligned} \tag{1.13}$$

Let the zeros of $p_n(x)$ be denoted by $x_{n,n} < \cdots < x_{n,2} < x_{n,1}$. Based on the differential equation (1.12), Chen and Ismail also showed that the largest zero $x_{n,1}$ has the asymptotic formula

$$x_{n,1} \approx X_n - i_1 \left(\frac{2a_n}{6A_n^2(2a_n)}\right)^{1/3}, \tag{1.14}$$

where i_1 is the first positive zero of Airy's function defined in [9, pp. 18 and 377] and

$$X_n = \sqrt{4a_n^2 + 2a_n/A_n(2a_n)}. \tag{1.15}$$

Moreover, they conjectured that the kth zero $x_{n,k}$ satisfies

$$x_{n,k} \approx X_n - i_k \left(\frac{2a_n}{6A_n^2(2a_n)} \right)^{1/3}, \tag{1.16}$$

where i_k is the kth positive zero of Airy's function.

In this paper, we shall extend the method used in [2] to derive an asymptotic formula for $p_n(x)$ with the weight function given by (1.1) and (1.8), which holds uniformly in an interval containing the critical value $x = 2a_n$. Our method is based on the turning-point theory developed in [7, Chap. 11]. Also, we shall construct a three-term asymptotic expansion for the zero $x_{n,k}$ for any fixed k. In particular, we shall establish the conjecture stated in (1.16).

2. Transformation to canonical form. In this section, we first show that the coefficient function $V(x,n)$ in (1.12) has turning points and then use the Liouville transformation to bring it to a canonical form. To this end, we note that the functions $A_n(x)$ and $B_n(x)$ in (1.6) and (1.7) satisfy the identity

$$B_n(x) + B_{n+1}(x) = \frac{x}{a_n} A_n(x) - v'(x); \tag{2.1}$$

see [3, (2.2)]. Differentiating both sides of (2.1) gives

$$B_n'(x) + B_{n+1}'(x) = \frac{A_n(x)}{a_n} + \frac{x}{a_n} A_n'(x) - v''(x). \tag{2.2}$$

By inserting (2.1) and (2.2) in (1.13), it can be verified that

$$\begin{aligned} V(x,n) = {} & \left[1 - \frac{x^2}{(2a_n)^2}\right] A_n^2(x) - A_n(x)\left[A_n(x) - A_{n-1}(x)\frac{a_n}{a_{n-1}}\right] \\ & - \frac{x}{2a_n} A_n(x)[B_n(x) - B_{n+1}(x)] + \frac{A_n(x)}{2a_n} \\ & + \frac{1}{2}[B_n'(x) - B_{n+1}'(x)] - \frac{1}{4}[B_n(x) - B_{n+1}(x)]^2 \\ & - \frac{1}{2}\frac{A_n'(x)}{A_n(x)}[B_n(x) - B_{n+1}(x)] + \frac{1}{2}\frac{A_n''(x)}{A_n(x)} - \frac{3}{4}\left(\frac{A_n'(x)}{A_n(x)}\right)^2. \end{aligned} \tag{2.3}$$

This equation gives the first indication that $V(x,n)$ has two turning points located at or near $x = \pm 2a_n$. Let

$$\lambda = 2a_n \qquad \text{and} \qquad x = \lambda w, \tag{2.4}$$

and define

$$W(\lambda, w) \equiv Y(\lambda w). \tag{2.5}$$

Equation (1.12) now becomes

$$\frac{d^2W}{dw^2} = U(\lambda w, n)W, \tag{2.6}$$

where

$$U(\lambda w, n) = -\lambda^2 V(\lambda w, n). \tag{2.7}$$

From (2.3), we have

$$\begin{aligned} U(\lambda w, n) = {} & \lambda^2(w^2-1)A_n^2(\lambda w) + \lambda^2 A_n(\lambda w)\left[A_n(\lambda w) - A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}}\right] \\ & + \lambda^2 A_n(\lambda w) w[B_n(\lambda w) - B_{n+1}(\lambda w)] - \lambda A_n(\lambda w) \\ & + \frac{\lambda^2}{4}[B_n(\lambda w) - B_{n+1}(\lambda w)]^2 - \frac{\lambda^2}{2}[B_n'(\lambda w) - B_{n+1}'(\lambda w)] \\ & + \frac{\lambda^2}{2}\frac{A_n'(\lambda w)}{A_n(\lambda w)}[B_n(\lambda w) - B_{n+1}(\lambda w)] - \frac{\lambda^2}{2}\frac{A_n''(\lambda w)}{A_n(\lambda w)} + \frac{3\lambda^2}{4}\left[\frac{A_n'(\lambda w)}{A_n(\lambda w)}\right]^2. \end{aligned} \tag{2.8}$$

To derive an asymptotic expansion for $U(\lambda w, n)$, as $\lambda \to \infty$, we first need to find the corresponding expansions for $A_n(\lambda w)$ and $B_n(\lambda w)$.

Put

$$c_{n,k,l} = \int_{-\infty}^{\infty} x^l p_n(x) p_{n-k}(x) w(x) dx, \qquad n, l \geq 0, \qquad n \geq k. \tag{2.9}$$

From the recurrence relation (1.10), it is readily verified by induction that p_n is an odd polynomial when n is odd and $p_n(x)$ is an even polynomial when n is even. Since $w(x)$ is also an even function by (1.8), it follows from (2.9) that

$$c_{n,k,l} = 0 \qquad \text{if} \quad k+l \quad \text{is odd.} \tag{2.10}$$

Applying (1.10) l times, we obtain

$$x^l p_n(x) = \sum_{k=-l}^{l} d_{n,k,l}\, p_{n-k}(x). \tag{2.11}$$

Substituting (2.11) in (2.9), we get by orthonormality

$$c_{n,k,l} = \begin{cases} d_{n,k,l}, & |k| \leq l, \\ 0, & |k| > l. \end{cases} \tag{2.12}$$

Hence (2.11) becomes

$$x^l p_n(x) = \sum_{k=-l}^{l} c_{n,k,l}\, p_{n-k}(x). \tag{2.13}$$

Coupling (1.10) and (2.9) gives

$$\begin{aligned} c_{n,k,l} &= \int_{-\infty}^{+\infty} x^{l-1} p_n(x)[a_{n-k+1}\, p_{n-k+1}(x) + a_{n-k}\, p_{n-k-1}(x)] w(x) dx \\ &= a_{n-k+1} \int_{-\infty}^{+\infty} x^{l-1} p_n(x)\, p_{n-(k-1)}(x) w(x) dx \\ &\quad + a_{n-k} \int_{-\infty}^{+\infty} x^{l-1} p_n(x)\, p_{n-(k+1)}(x) w(x) dx \\ &= a_{n-k+1}\, c_{n,k-1,l-1} + a_{n-k}\, c_{n,k+1,l-1}. \end{aligned}$$

So we also have the recurrence relation

$$c_{n,k,l} = a_{n-k}\, c_{n,k+1,l-1} + a_{n-k+1}\, c_{n,k-1,l-1}. \tag{2.14}$$

LEMMA 2.1. (i) *When $k+l$ is odd or $|k| > l$, we have $c_{n,k,l} = 0$.* (ii) *When $k+l$ is even and $|k| \le l, c_{n,k,l}$ is a sum of (monic) monomials of the form $a_{n-j_1}a_{n-j_2}\cdots a_{n-j_l}$; that is,*

$$c_{n,k,l} = \sum_{j_1,j_2,\ldots,j_l} a_{n-j_1}a_{n-j_2}\cdots a_{n-j_l}, \tag{2.15}$$

where $k-l \le j_i \le k+l-1$ and $1 \le i \le l$. (iii) *There are $\binom{l}{\frac{1}{2}(k+l)}$ nonzero terms in the summation on the right-hand side of* (2.15). (iv) *The sum of all indices $j_1,\ldots,j_l$ of nonzero terms in* (2.15) *is equal to*

$$s_{k,l} = \sum_{j=0}^{l-1}\sum_{s=0}^{j}\binom{j}{s}\left[(k+j-2s)\binom{l-j}{\frac{1}{2}(k+l)-s} - \binom{l-j-1}{\frac{1}{2}(k+l)-s-1}\right]. \tag{2.16}$$

Proof. Statement (i) is given in (2.10) and (2.12). Statement (ii) can be shown by repeated application of (2.14) l times. Note that each time when we apply (2.14), the k-index in the first term on the right-hand side goes up by one, whereas the k-index in the second term goes down by one. Furthermore, when $|k| > l$, we have $c_{n,k,l} = 0$. Statement (iii) is proved by induction. When $l = 1$, we have $k = \pm 1$ since $|k| \le l$ and $k+l$ is even by hypothesis. From (2.14), we have $c_{n,1,1} = a_n$ and $c_{n,-1,1} = a_{n+1}$. Hence, in either case, there is only one term on the right-hand side of (2.15); i.e., statement (iii) holds when $l = 1$. Assume that it holds for $l = \tilde{l}-1$. Then $c_{n,k,\tilde{l}-1}$ has $\binom{\tilde{l}-1}{\frac{1}{2}(k+\tilde{l}-1)}$ nonzero terms whenever $k+\tilde{l}-1$ is even. Thus, by (2.14), when $k+\tilde{l}$ is even, $c_{n,k,\tilde{l}}$ has $\binom{\tilde{l}-1}{\frac{1}{2}(k+\tilde{l})} + \binom{\tilde{l}-1}{\frac{1}{2}(k+\tilde{l}-2)}$ nonzero terms. In view of the identity

$$\binom{N}{M} = \binom{N-1}{M} + \binom{N-1}{M-1}, \tag{2.17}$$

it follows that statement (iii) also holds for $l = \tilde{l}$. To prove statement (iv), we again use induction. By (iii), $s_{1,1} = 0$ and $s_{-1,1} = -1$. Direct calculation shows that the right-hand side of (2.16) is also equal to 0 when $l = 1$ and $k = 1$, and -1 when $l = 1$ and $k = -1$, thus establishing (iv) in this case. Next, assume that (2.16) holds for $l = \tilde{l}-1$. Then k must satisfy $-\tilde{l}+1 \le k \le \tilde{l}-1$ and $k+\tilde{l}-1$ must be even. Consider the case $l = \tilde{l}$, $-\tilde{l} \le k \le \tilde{l}$ and $k+\tilde{l}$ even. From (2.14), one can see that

$$s_{k,\tilde{l}} = k\binom{\tilde{l}-1}{\frac{1}{2}(k+\tilde{l})} + s_{k+1,\tilde{l}-1} + (k-1)\binom{\tilde{l}-1}{\frac{1}{2}(k+\tilde{l})-1} + s_{k-1,\tilde{l}-1} \tag{2.18}$$

for $-\tilde{l} \le k \le \tilde{l}$ and $k+\tilde{l}$ even; here we have assumed that $s_{k,l} = 0$ if $|k| > l$. In particular, when $k = \pm\tilde{l}$,

$$s_{\tilde{l},\tilde{l}} = \tilde{l} - 1 + s_{\tilde{l}-1,\tilde{l}-1}, \qquad s_{-\tilde{l},\tilde{l}} = -\tilde{l} + s_{-\tilde{l}+1,\tilde{l}-1}. \tag{2.19}$$

From (2.18), we have by the induction hypothesis

$$s_{k,\tilde{l}} = k\binom{\tilde{l}}{\frac{1}{2}(k+\tilde{l})} - \binom{\tilde{l}-1}{\frac{1}{2}(k+\tilde{l})-1}$$

$$\tag{2.20} +\sum_{j=0}^{\tilde{l}-2}\sum_{s=0}^{j}\binom{j}{s}\left[(k+j-2s+1)\binom{\tilde{l}-j-1}{\frac{1}{2}(k+\tilde{l})-s}-\binom{\tilde{l}-j-2}{\frac{1}{2}(k+\tilde{l})-s-1}\right]$$

$$+\sum_{j=0}^{\tilde{l}-2}\sum_{s=0}^{j}\binom{j}{s}\left[(k+j-2s-1)\binom{\tilde{l}-j-1}{\frac{1}{2}(k+\tilde{l})-s-1}-\binom{\tilde{l}-j-2}{\frac{1}{2}(k+\tilde{l})-s-2}\right].$$

After reindexing the last term on the right-hand side of (2.20) and applying (2.17), it can be shown that

$$\tag{2.21} s_{k,\tilde{l}}=\sum_{j=0}^{\tilde{l}-1}\sum_{s=0}^{j}\binom{j}{s}\left[(k+j-2s)\binom{\tilde{l}-j}{\frac{1}{2}(k+\tilde{l})-s}-\binom{\tilde{l}-j-1}{\frac{1}{2}(k+\tilde{l})-s-1}\right],$$

which is exactly (2.16). The first two terms on the right-hand side of (2.20) are included in the last sum with $j=0$. When $k=\pm l$, (2.16) reduces

$$s_{\tilde{l},\tilde{l}}=\sum_{j=0}^{\tilde{l}-1}(\tilde{l}-1-j) \qquad \text{and} \qquad s_{-\tilde{l},\tilde{l}}=\sum_{j=0}^{\tilde{l}-1}(\tilde{l}-j),$$

which agree with what can be obtained from (2.19) directly, thus proving the case $l=\tilde{l}$. This completes the proof of the lemma. □

We now proceed to find the asymptotic behavior of $A_n(x)$ and $B_n(x)$. Substituting (1.9) into (1.6) and (1.7), we get

$$A_n(x)=2a_n\sum_{k=1}^{m}kv_k\sum_{l=0}^{2k-2}x^{2k-l-2}c_{n,0,l}$$

and

$$B_n(x)=2a_n\sum_{k=1}^{m}kv_k\sum_{l=0}^{2k-2}x^{2k-l-2}c_{n,1,l}.$$

In view of (2.10), the last two equations can be written as

$$\tag{2.22} A_n(x)=2a_n\sum_{k=0}^{m-1}(k+1)v_{k+1}\sum_{l=0}^{k}x^{2k-2l}c_{n,0,2l}$$

and

$$\tag{2.23} B_n(x)=2a_n\sum_{k=1}^{m-1}(k+1)v_{k+1}\sum_{l=0}^{k-1}x^{2k-2l-1}c_{n,1,2l+1}.$$

In [1], Bauldry, Máté, and Nevai proved that for any $N>0$, there exist constants $\eta_1,\ldots,\eta_N$ such that

$$\tag{2.24} a_n=(L_m n)^{1/2m}\left[1+\frac{\eta_1}{n^{1/m}}+\frac{\eta_2}{n^{2/m}}+\cdots+\frac{\eta_N}{n^{N/m}}+O\left(\frac{1}{n^{(N+1)/m}}\right)\right]$$

as $n \to \infty$, where

$$L_m = \frac{m!(m-1)!}{(2m)!}. \tag{2.25}$$

These coefficients $\eta_1, \eta_2, \ldots$ will be determined later in section 3. In the special case $v(x) = x^{2m}$, Máté, Nevai, and Zaslavsky [5] earlier have given the result

$$a_n = (L_m n)^{1/2m}\left[1 + \frac{\eta}{n^2} + O\left(\frac{1}{n^4}\right)\right] \qquad \text{as } n \to \infty, \tag{2.26}$$

where η is some constant independent of n. Since $\lambda = 2a_n$ by (2.4), we also have from (2.24) $\lambda = 2(L_m n)^{1/2m}[1 + O(n^{-1/m})]$ or, equivalently,

$$n = \frac{\lambda^{2m}}{2^{2m}L_m}\left[1 + O\left(\frac{1}{\lambda^2}\right)\right]. \tag{2.27}$$

Note that $n^{-1/m} = O(\lambda^{-2})$. Taking $N = m$ in (2.24) gives

$$a_{n-j} = [L_m(n-j)]^{1/2m}\left[1 + \frac{\eta_1}{(n-j)^{1/m}} + \cdots + \frac{\eta_m}{(n-j)} + O\left(\frac{1}{n^{1+1/m}}\right)\right]$$

for any integer j. Upon simplification, we get

$$a_{n-j} = (L_m n)^{1/2m}\left[1 - \frac{j}{2mn} + O\left(\frac{1}{n^2}\right)\right]\left[1 + \frac{\eta_1}{n^{1/m}} + \cdots + \frac{\eta_m}{n} + O\left(\frac{1}{n^{1+1/m}}\right)\right]. \tag{2.28}$$

Comparing (2.28) with (2.24) yields

$$a_{n-j} = a_n\left[1 - \frac{j}{2mn} + O\left(\frac{1}{n^{1+1/m}}\right)\right]. \tag{2.29}$$

Substituting (2.29) in (2.15), we have

$$c_{n,k,l} = \binom{l}{\frac{1}{2}(k+l)} a_n^l \left[1 + \frac{\tilde{c}_{k,l}}{2mn} + O\left(\frac{1}{n^{1+1/m}}\right)\right], \tag{2.30}$$

where $\tilde{c}_{k,l} = -s_{k,l}/\binom{l}{\frac{1}{2}(k+l)}$, $s_{k,l}$ being defined in (2.16). Inserting (2.30) into (2.22) and (2.23), we obtain

$$\begin{aligned} A_n(x) = 2a_n \sum_{k=0}^{m-1} & (k+1)v_{k+1} \\ & \cdot \sum_{l=0}^{k} x^{2k-2l} a_n^{2l} \binom{2l}{l} \left[1 + \frac{\tilde{c}_{0,2l}}{2mn} + O\left(\frac{1}{n^{1+1/m}}\right)\right] \end{aligned} \tag{2.31}$$

and

$$\begin{aligned} B_n(x) = 2a_n \sum_{k=1}^{m-1} & (k+1)v_{k+1} \\ & \cdot \sum_{l=0}^{k-1} x^{2k-2l-1} a_n^{2l+1} \binom{2l+1}{l+1} \left[1 + \frac{\tilde{c}_{1,2l+1}}{2mn} + O\left(\frac{1}{n^{1+1/m}}\right)\right]. \end{aligned} \tag{2.32}$$

Put $x = \lambda w$. Since $2a_n = \lambda$, it follows from (2.27) that

$$\begin{aligned} A_n(\lambda w) = \lambda \sum_{k=0}^{m-1} & (k+1)v_{k+1}\lambda^{2k} \\ & \cdot \sum_{l=0}^{k} w^{2k-2l}2^{-2l}\binom{2l}{l}\left[1 + \frac{\tilde{d}_{2l}}{\lambda^{2m}} + O\left(\frac{1}{\lambda^{2m+2}}\right)\right] \end{aligned} \tag{2.33}$$

and

$$\begin{aligned} B_n(\lambda w) = \lambda \sum_{k=1}^{m-1} & (k+1)v_{k+1}\lambda^{2k} \\ & \cdot \sum_{l=0}^{k-1} w^{2k-2l-1}2^{-2l-1}\binom{2l+1}{l+1}\left[1 + \frac{\tilde{d}_{2l+1}}{\lambda^{2m}} + O\left(\frac{1}{\lambda^{2m+2}}\right)\right], \end{aligned} \tag{2.34}$$

where $\tilde{d}_{2l} = 2^{2m}L_m\tilde{c}_{0,2l}/2m$ and $\tilde{d}_{2l+1} = 2^{2m}L_m\tilde{c}_{1,2l+1}/2m$. Note that $A_n(\lambda w)$ and $B_n(\lambda w)$ are polynomials in w of degree $2m-2$ and $2m-3$, respectively. Formulas (2.33) and (2.34) clearly imply that

$$\begin{aligned} A_n(\lambda w) &= \sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k+1}\sum_{l=0}^{k} w^{2k-2l}2^{-2l}\binom{2l}{l} + O\left(\frac{1}{\lambda}\right) \\ &\sim \lambda^{2m-1}mv_m\sum_{l=0}^{m-1} w^{2m-2l-2}2^{-2l}\binom{2l}{l} \end{aligned} \tag{2.35}$$

and

$$\begin{aligned} B_n(\lambda w) &= \sum_{k=1}^{m-1}(k+1)v_{k+1}\lambda^{2k+1}\sum_{l=0}^{k-1} w^{2k-2l-1}2^{-2l-1}\binom{2l+1}{l+1} + O\left(\frac{1}{\lambda}\right) \\ &\sim \lambda^{2m-1}mv_m\sum_{l=0}^{m-2} w^{2m-2l-3}2^{-2l-1}\binom{2l+1}{l+1}. \end{aligned} \tag{2.36}$$

Since $A_n(\lambda w)$ and $B_n(\lambda w)$ are polynomials in w, the last two asymptotic formulas hold uniformly with respect to w in any bounded interval. In fact, from here on, all O-symbols will be used to mean that they are uniform with respect to w in any bounded interval, except for cases otherwise indicated.

In (2.31), we now replace n by $n-1$. Coupling the resulting formula with (2.29), we obtain

$$\begin{aligned} A_{n-1}(x)\frac{a_n}{a_{n-1}} = 2a_n & \sum_{k=0}^{m-1}(k+1)v_{k+1} \\ & \cdot \sum_{l=0}^{k} w^{2k-2l}a_n^{2l}\binom{2l}{l}\left[1 + \frac{\tilde{c}_{0,2l} - 2l}{2mn} + O\left(\frac{1}{n^{1+1/m}}\right)\right]. \end{aligned} \tag{2.37}$$

In terms of λ, (2.37) can be written as

$$\begin{aligned} A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}} = \lambda & \sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k} \\ & \cdot \sum_{l=0}^{k} w^{2k-2l}2^{-2l}\binom{2l}{l}\left[1 + \frac{\tilde{d}'_{2l}}{\lambda^{2m}} + O\left(\frac{1}{\lambda^{2m+2}}\right)\right], \end{aligned}$$

where $\tilde{d}'_{2l} = \tilde{d}_{2l} - \frac{l}{m}2^{2m}L_m$. Hence

$$A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}} = \sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k+1}\sum_{l=0}^{k} w^{2k-2l}2^{-2l}\binom{2l}{l} + O\left(\frac{1}{\lambda}\right). \tag{2.38}$$

A comparison of (2.35) and (2.38) gives

$$A_n(\lambda w) - A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}} = O\left(\frac{1}{\lambda}\right). \tag{2.39}$$

In a similar manner, it can be shown that

$$B_n(\lambda w) - B_{n+1}(\lambda w) = O\left(\frac{1}{\lambda}\right). \tag{2.40}$$

By differentiating (2.22) and using (2.30), we have

$$\begin{aligned} A'_n(\lambda w) = \sum_{k=1}^{m-1}&(k+1)v_{k+1}\lambda^{2k} \\ &\cdot \sum_{l=0}^{k-1} 2(k-l)w^{2k-2l-1}2^{-2l}\binom{2l}{l}\left[1 + \frac{\tilde{d}_{2l}}{\lambda^{2m}} + O\left(\frac{1}{\lambda^{2m+2}}\right)\right], \end{aligned}$$

from which it follows that

$$A'_n(\lambda w) = O(\lambda^{2m-2}). \tag{2.41}$$

The same analysis yields

$$B'_n(\lambda w) = O(\lambda^{2m-2}), \tag{2.42}$$

$$A''_n(\lambda w) = O(\lambda^{2m-3}), \tag{2.43}$$

and

$$B'_n(\lambda w) - B'_{n+1}(\lambda w) = O\left(\frac{1}{\lambda^2}\right). \tag{2.44}$$

We now return to the function $U(\lambda w, n)$ in (2.8) and write

$$U(\lambda w, n) = U_0(\lambda w, n) + U_1(\lambda w, n) + U_2(\lambda w, n), \tag{2.45}$$

where

$$U_0(\lambda w, n) = \lambda^2(w^2-1)A_n^2(\lambda w), \tag{2.46}$$

$$\begin{aligned} U_1(\lambda w, n) = \lambda^2 A_n(\lambda w)&\left[A_n(\lambda w) - A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}}\right] \\ &+ \lambda^2 A_n(\lambda w)w[B_n(\lambda w) - B_{n+1}(\lambda w)] - \lambda A_n(\lambda w), \end{aligned} \tag{2.47}$$

and

$$\begin{aligned} U_2(\lambda w, n) = \frac{\lambda^2}{4}&[B_n(\lambda w) - B_{n+1}(\lambda w)]^2 - \frac{\lambda^2}{2}[B'_n(\lambda w) - B'_{n+1}(\lambda w)] \\ &+ \frac{\lambda^2}{2}\frac{A'_n(\lambda w)}{A_n(\lambda w)}[B_n(\lambda w) - B_{n+1}(\lambda w)] - \frac{\lambda^2}{2}\frac{A''_n(\lambda w)}{A_n(\lambda w)} + \frac{3\lambda^2}{4}\left[\frac{A'_n(\lambda w)}{A_n(\lambda w)}\right]^2. \end{aligned} \tag{2.48}$$

From the asymptotic results in (2.35), (2.39)–(2.41), (2.43), and (2.44), it follows that

$$U_0(\lambda w, n) = O(\lambda^{4m}), \tag{2.49}$$

$$U_1(\lambda w, n) = O(\lambda^{2m}), \tag{2.50}$$

and

$$U_2(\lambda w, n) = O(1). \tag{2.51}$$

With these order estimates, the differential equation in (2.6) can be expressed as

$$\frac{d^2}{dw^2} W(\lambda, w) = [\lambda^{4m} H_0(\lambda, w) + \lambda^{2m} H_1(\lambda, w) + H_2(\lambda, w)] W(\lambda, w), \tag{2.52}$$

where $H_0(\lambda, w) = \lambda^{-4m} U_0(\lambda w, n), H_1(\lambda, w) = \lambda^{-2m} U_1(\lambda w, n)$, and $H_2(\lambda, w) = U_2(\lambda w, n)$. From the definitions of $H_0(\lambda, w)$ and $H_1(\lambda, w)$, it is clear that they are polynomials in w. Since $A_n(\lambda w) \neq 0, H_2(\lambda, w)$ is a real analytic function of w. Note that for each $i = 0, 1, 2$, we have $H_i(\lambda, w) = O(1)$. Since

$$H_0(\lambda, w) = \lambda^{-4m+2}(w^2 - 1) A_n^2(\lambda w) \tag{2.53}$$

vanishes only at $w = \pm 1$, (2.52) has two turning points, one at $w = 1$ and the other at $w = -1$. From the recurrence relation (1.10), it can be readily verified by induction that $p_n(x)$ satisfies the reflection formula

$$p_n(x) = (-1)^n p_n(-x). \tag{2.54}$$

Thus we need consider only the turning point $w = +1$.

2.1. Liouville transformation. Following [7, p. 398], we make the Liouville transformation

$$\zeta \left(\frac{d\zeta}{dw} \right)^2 = H_0(\lambda, w), \qquad Z = \left(\frac{d\zeta}{dw} \right)^{1/2} W. \tag{2.55}$$

Integration of the first equation in (2.55) gives

$$\zeta(\lambda, w) = \begin{cases} \left[\dfrac{3}{2} \dfrac{1}{\lambda^{2m-1}} \displaystyle\int_1^w \sqrt{t^2 - 1}\, A_n(\lambda t) dt \right]^{2/3}, & w \geq 1, \\ -\left[\dfrac{3}{2} \dfrac{1}{\lambda^{2m-1}} \displaystyle\int_w^1 \sqrt{1 - t^2}\, A_n(\lambda t) dt \right]^{2/3}, & |w| \leq 1. \end{cases} \tag{2.56}$$

Setting

$$\hat{H}_0(\lambda, w) = \frac{H_0(\lambda, w)}{\zeta}, \tag{2.57}$$

we have from (2.55)

$$Z = \hat{H}_0^{1/4}(\lambda, w) W. \tag{2.58}$$

Note that $d\zeta/dw = \hat{H}_0^{1/2}(\lambda, w) > 0$ for $w \neq 1$, and hence that the relation in (2.56) defines a one-to-one correspondence between w and ζ. Since (2.35) holds uniformly for w in bounded intervals, we have

$$\zeta(\lambda, w) \to \zeta_\infty(w) \tag{2.59}$$

uniformly for bounded w, where

$$\zeta_\infty(w) = \begin{cases} \left[\frac{3}{2}mv_m \sum_{l=0}^{m-1} 2^{-2l}\binom{2l}{l}\int_1^w \sqrt{t^2-1}\, t^{2m-2l-2}dt\right]^{2/3}, & w \geq 1, \\ -\left[\frac{3}{2}mv_m \sum_{l=0}^{m-1} 2^{-2l}\binom{2l}{l}\int_w^1 \sqrt{1-t^2}\, t^{2m-2l-2}dt\right]^{2/3}, & |w| \leq 1. \end{cases}$$

(2.60)

Put $u = \lambda^{2m}$. The Liouville transformation (2.55) then takes (2.52) into the form

$$\frac{d^2Z}{d\zeta^2} = \{u^2\zeta + u\phi(\lambda,\zeta) + \psi(\lambda,\zeta)\}Z, \tag{2.61}$$

where

$$\phi(\lambda,\zeta) = \frac{H_1(\lambda,w)}{\hat{H}_0(\lambda,w)} = \frac{H_1(\lambda,w)}{H_0(\lambda,w)}\zeta \tag{2.62}$$

and

$$\psi(\lambda,\zeta) = \frac{H_2(\lambda,w)}{\hat{H}_0(\lambda,w)} - \frac{1}{\hat{H}_0(\lambda,w)^{3/4}}\frac{d^2}{dw^2}\left(\frac{1}{\hat{H}_0(\lambda,w)^{1/4}}\right). \tag{2.63}$$

The following result is an analogue of Lemma 3.1 in [7, p. 399]. If $f(\lambda,x)$ represents a function of x with a parameter λ, it will be understood that $f^{(k)}(\lambda,x)$ denotes the kth derivative of f with respect to x.

Lemma 2.2. *In a given interval (a,b), let $f(\lambda,x)$ be an n-times continuously differentiable function of x, and let $x_0 \in (a,b)$. Define*

$$g(\lambda,x) = \begin{cases} \frac{1}{(x-x_0)^{3/2}}\int_{x_0}^x (t-x_0)^{1/2}f(\lambda,t)dt, & x \geq x_0, \\ \frac{1}{(x_0-x)^{3/2}}\int_x^{x_0} (x_0-t)^{1/2}f(\lambda,t)dt, & x < x_0. \end{cases} \tag{2.64}$$

Then $g(\lambda,x)$ is n-times continuously differentiable with respect to x. Moreover, if $f(\lambda,x)$ and its derivatives $f^{(k)}(\lambda,x)$, respectively, tend to $f_\infty(x)$ and its derivatives $f_\infty^{(k)}(x)$ uniformly in a closed subinterval of (a,b), then $g(\lambda,x)$ and its derivatives $g^{(k)}(\lambda,x)$ also, respectively, tend to

$$g_\infty(x) = \begin{cases} \frac{1}{(x-x_0)^{3/2}}\int_{x_0}^x (t-x_0)^{1/2}f_\infty(t)dt, & x \geq x_0, \\ \frac{1}{(x_0-x)^{3/2}}\int_x^{x_0} (x_0-t)^{1/2}f_\infty(t)dt, & x < x_0, \end{cases} \tag{2.65}$$

and its derivatives $g_\infty^{(k)}(x)$ uniformly in this closed subinterval.

Proof. Since the proof of this result is similar to that of Lemma 3.1 in [7, p. 399], we present here only a brief outline of the argument. Consider $x \in [x_0, b)$. By repeated use of the mean-value theorem and repeated integration by parts, it can be shown that

$$g^{(k)}(\lambda,x) = \frac{1}{(x-x_0)^{k+3/2}}\int_{x_0}^x (t-x_0)^{k+1/2}f^{(k)}(\lambda,t)dt, \qquad x > x_0,$$

and

$$g^{(k)}(\lambda, x) \longrightarrow \frac{1}{k+\frac{3}{2}} f^{(k)}(\lambda, x_0) \qquad\qquad \text{as } x \to x_0$$

for $k = 0, 1, \ldots, n$. These two results also hold when $\lambda = \infty$. This establishes the first part of the lemma. If for any $\varepsilon > 0$ we have $|f^{(k)}(\lambda, x) - f_\infty^{(k)}(x)| < \varepsilon$ uniformly for x in a closed subinterval of (a, b), then it can be verified that

$$|g^{(k)}(\lambda, x) - g_\infty^{(k)}(x)| < \frac{\varepsilon}{k+\frac{3}{2}}$$

also uniformly for x in that subinterval, thus proving the second part of the lemma. □

To apply the above result, we replace the variable x by w and take $(a, b) = (-1, \infty)$, $x_0 = 1$, and

$$f(\lambda, w) = \left[\frac{H_0(\lambda, w)}{w-1}\right]^{1/2} = \frac{1}{\lambda^{2m-1}}(w+1)^{1/2} A_n(\lambda w). \tag{2.66}$$

Let $g(\lambda, w)$ be defined as in (2.64); i.e.,

$$g(\lambda, w) = \begin{cases} \dfrac{1}{(v-1)^{3/2}} \displaystyle\int_1^w (v-1)^{1/2} f(\lambda, v)dv, & w \geq 1, \\ \dfrac{1}{(1-w)^{3/2}} \displaystyle\int_w^1 (1-v)^{1/2} f(\lambda, v)dv, & |w| \leq 1. \end{cases} \tag{2.67}$$

By Lemma 2.2, $g(\lambda, w)$ is n-times continuously differentiable in $(-1, \infty)$. Moreover, the limits of $g(\lambda, w)$ and its derivatives $g^{(k)}(\lambda, w)$, as $\lambda \to \infty$, exist uniformly on $[-1+\varepsilon, M]$ for any $0 < \varepsilon \ll 1$ and $1 \ll M < \infty$. Note that $g(\lambda, w) > 0$ since $f(\lambda, w) > 0$, and that from (2.56) and (2.67) we have

$$\frac{\zeta(\lambda, w)}{w-1} = \left[\frac{3}{2} g(\lambda, w)\right]^{2/3}.$$

Hence, it follows that $\zeta(\lambda, w)/(w-1)$ is nonvanishing, and that $\zeta(\lambda, w)$ is n-times continuously differentiable. Furthermore, the limits of the derivatives $\zeta^{(k)}(\lambda, w)$, as $\lambda \to \infty$, exist uniformly on $[-1+\varepsilon, M]$. In particular,

$$\zeta'(\lambda, 1) = \lim_{w \to 1} \frac{\zeta(\lambda, w)}{w-1} = \left[\frac{3}{2} g(\lambda, 1)\right]^{2/3} = 2^{1/3} \left[\frac{A_n(\lambda)}{\lambda^{2m-1}}\right]^{2/3} > 0$$

for all λ. When $w \neq 1$, we have already seen that $\zeta'(\lambda, w) = [\hat{H}_0(\lambda, w)]^{1/2} > 0$. Therefore, $\zeta'(\lambda, w)$ is a positive and differentiable function in $(-1, \infty)$. By (2.57), $\hat{H}_0(\lambda, w)$ is also n-times continuously differentiable and has a uniform limit, as $\lambda \to \infty$, on closed subintervals of $(-1, \infty)$. For sufficiently large λ, $\hat{H}_0(\lambda, w)$ is strictly positive in $[-1+\varepsilon, M]$. By using the chain rule and the inverse function theorem, it is evident from (2.62) and (2.63) that $\phi(\lambda, \zeta)$ and $\psi(\lambda, \zeta)$ are n-times continuously differentiable with respect to ζ. By (2.59), $\zeta(\lambda, -1+\varepsilon) \to \zeta_\infty(-1+\varepsilon)$ and $\zeta(\lambda, M) \to \zeta_\infty(M)$ as $\lambda \to \infty$, and by (2.60) we have $\zeta_\infty(-1+\varepsilon) < 0$ and $\zeta_\infty(M) > 0$. The limits of $\phi(\lambda, \zeta), \psi(\lambda, \zeta)$ and all their derivatives exist uniformly on the closed interval $[\zeta_\infty(-1+\varepsilon), \zeta_\infty(M)]$.

Motivated by (2.13) in [2], we define

$$\Phi(\zeta) = \Phi(\lambda, \zeta) = \begin{cases} \dfrac{1}{2\zeta^{1/2}} \displaystyle\int_0^\zeta \frac{\phi(\lambda, v)}{v^{1/2}} dv, & \zeta > 0, \\ \dfrac{1}{2(-\zeta)^{1/2}} \displaystyle\int_\zeta^0 \frac{\phi(\lambda, v)}{(-v)^{1/2}} dv, & \zeta < 0. \end{cases} \tag{2.68}$$

As in the proof of Lemma 2.2, it can be shown that $\Phi(\zeta)$ is continuous at $\zeta = 0$ and

$$\Phi(0) = \phi(\lambda, 0).$$

Integration by parts gives

$$\Phi'(\lambda, \zeta) = \frac{\pm 1}{2|\zeta|^{3/2}} \int_0^\zeta |v|^{1/2} \phi'(\lambda, v) dv,$$

where the ± signs depend on $\zeta > 0$ or $\zeta < 0$, and by the mean-value theorem, $\Phi'(\lambda, \zeta)$ is continuous at $\zeta = 0$. From Lemma 2.2, it follows that $\Phi(\lambda, \zeta)$ has n-times continuous derivatives and that all its derivatives tend to their limits, as $\lambda \to \infty$, uniformly on $[\zeta_\infty(-1+\varepsilon), \zeta_\infty(M)]$. As a consequence, $\Phi(\zeta)$ and its derivatives $\Phi^{(k)}(\zeta), k = 1, \ldots, n$, are uniformly bounded on $[\zeta_\infty(-1+\varepsilon), \zeta_\infty(M)]$ for all sufficiently large λ.

2.2. Asymptotic solutions. In the following, we wish to present two linearly independent asymptotic solutions to (2.61). This result is analogous to Theorem 1 in [2] or Theorem 3.1 in [7, p. 399]. Before stating the theorem, we first recall the modulus function $M(x)$ and the weight function $E(x)$ associated with the Airy functions $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$; cf. [7, p. 395]. Let c denote the negative root of the equation

$$\mathrm{Ai}(x) = \mathrm{Bi}(x)$$

of smallest absolute value. Define $E(x) = 1$ for $-\infty < x \le c$,

$$E(x) = \{\mathrm{Bi}(x)/\mathrm{Ai}(x)\}^{1/2}, \qquad c \le x < \infty,$$

and

$$M(x) = \{E^2(x)\,\mathrm{Ai}^2(x) + E^{-2}(x)\,\mathrm{Bi}^2(x)\}^{1/2},$$

where $E^{-1}(x) = 1/E(x)$. The phase function $\theta(x)$ is defined by

$$E(x)\,\mathrm{Ai}(x) = M(x)\sin\theta(x), \qquad E^{-1}(x)\,\mathrm{Bi}(x) = M(x)\cos\theta(x),$$

or, equivalently,

$$\theta(x) = \tan^{-1}\{E^2(x)\,\mathrm{Ai}(x)/\mathrm{Bi}(x)\}.$$

Modulus and phase functions are also needed for the derivatives of the Airy functions, and they are defined by

$$E(x)\,\mathrm{Ai}'(x) = N(x)\sin\omega(x), \qquad E^{-1}(x)\,\mathrm{Bi}'(x) = N(x)\cos\omega(x).$$

For convenience, we introduce the function (cf. [7, p. 429])

$$G(\zeta) = G(\lambda, \zeta) = \zeta\Phi'^2 + 2\Phi\Phi' + \frac{\Phi\Phi'^2}{u} + \frac{3\Phi''^2 - 2\Phi'\Phi''' - 2u\Phi'''}{4u^2(1 + \Phi'/u)^2}, \tag{2.69}$$

where Φ is given in (2.68) and $u = \lambda^{2m}$. The error control function is then defined by

$$\Psi(\zeta) = \Psi(\lambda,\zeta) = \int_0^\zeta \frac{\psi(v) - G(v)}{1 + \Phi'(v)/u}\left\{v + \frac{\Phi(v)}{u}\right\}^{-\frac{1}{2}} dv, \tag{2.70}$$

where $\psi(v) = \psi(v,\lambda)$ is given in (2.63); cf. [7, p. 399] and [2, (2.18)]. Clearly the integral in (2.70) is convergent for sufficiently large values of u and, as $\lambda \to \infty$, $\Psi(\lambda,\zeta)$ tends to a limiting function $\Psi_\infty(\zeta)$ uniformly on the interval $[\zeta_\infty(-1+\varepsilon), \zeta_\infty(M)]$ containing the origin $\zeta = 0$.

Here and thereafter, we shall often use f or $f(\zeta)$ to represent $f(\lambda,\zeta)$. For instance, both Φ and $\Phi(\zeta)$ will mean the function $\Phi(\lambda,\zeta)$ defined in (2.68).

THEOREM 2.3. *Equation* (2.61) *has a pair of twice continuously differentiable solutions $Z_1(\lambda,\zeta)$ and $Z_2(\lambda,\zeta)$, given by*

$$Z_1(\lambda,\zeta) = \left(1 + \frac{\Phi'(\zeta)}{u}\right)^{-\frac{1}{2}}\left\{\mathrm{Ai}\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) + \varepsilon_1(\lambda,\zeta)\right\}, \tag{2.71}$$

$$Z_2(\lambda,\zeta) = \left(1 + \frac{\Phi'(\zeta)}{u}\right)^{-\frac{1}{2}}\left\{\mathrm{Bi}\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) + \varepsilon_2(\lambda,\zeta)\right\}. \tag{2.72}$$

For sufficiently large λ, the error terms satisfy

$$\begin{aligned} &|\varepsilon_1(\lambda,\zeta)| \Big/ M\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right),\ |\varepsilon_1'(\lambda,\zeta)| \Big/ \left(u^{2/3} + \frac{\Phi'(\zeta)}{u^{1/3}}\right) N\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) \\ &\quad \le \frac{K\pi}{u} E^{-1}\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right)\mathcal{V}_{\zeta,\zeta_\infty(M)}(\Psi(\xi)) \exp\left\{\frac{K_0}{u}\mathcal{V}_{\zeta,\zeta_\infty(M)}(\Psi(\xi))\right\} \end{aligned} \tag{2.73}$$

and

$$\begin{aligned} &|\varepsilon_2(\lambda,\zeta)| \Big/ M\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right),\ |\varepsilon_2'(\lambda,\zeta)| \Big/ \left(u^{2/3} + \frac{\Phi'(\zeta)}{u^{1/3}}\right) N\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) \\ &\quad \le \frac{K\pi}{u} E\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right)\mathcal{V}_{\zeta_\infty(-1+\varepsilon),\zeta}(\Psi(\xi)) \exp\left\{\frac{K_0}{u}\mathcal{V}_{\zeta_\infty(-1+\varepsilon),\zeta}(\Psi(\xi))\right\}, \end{aligned} \tag{2.74}$$

where K, K_0 are positive constants, $\mathcal{V}_{a,b}(f)$ denotes the total variation of a function $f(\zeta)$ on an interval (a,b), and $\zeta_\infty(w)$ is the function given in (2.59)–(2.60).

Since the argument used in the proof of this theorem is along the same line as that for Theorem 3.1 in [7, p. 399], it will not be included in this paper.

From (2.73), we have

$$\varepsilon_1(\lambda,\zeta) = E^{-1}\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) M\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) O(u^{-1})$$

$$\varepsilon_1'(\lambda,\zeta) = E^{-1}\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) N\left(u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}}\right) O(u^{-1/3}).$$

In view of the asymptotic results [7, pp. 395–396]

$$E(x) \sim 2^{1/2}\exp\left(\frac{2}{3}x^{3/2}\right), \qquad M(x) \sim \pi^{-1/2}x^{-1/4} \qquad (x \to +\infty),$$

$$M(x) \sim \pi^{-1/2}(-x)^{-1/4} \qquad (x \to -\infty),$$

and

$$N(x) \sim \pi^{-1/2}|x|^{1/4} \qquad (x \to \pm\infty),$$

it follows that

$$\varepsilon_1(\lambda,\zeta) = O(u^{-1}), \qquad \varepsilon_1'(\lambda,\zeta) = O(u^{-1/6}) \tag{2.75}$$

as $\lambda \to \infty$, uniformly for ζ in $[\zeta_\infty(-1+\varepsilon), \zeta_\infty(M)]$. Moreover, if $\zeta \in [\zeta_\infty(-1+\varepsilon), \zeta_\infty(1-\varepsilon)]$, then

$$\varepsilon_1(\lambda,\zeta) = O(u^{-7/6}), \qquad \varepsilon_1'(\lambda,\zeta) = O(u^{-1/6}). \tag{2.76}$$

(Recall that $\zeta_\infty(1-\varepsilon)$ is negative.) If $\zeta \in [\zeta_\infty(1+\varepsilon), \zeta_\infty(M)]$, then both $\varepsilon_1(\lambda,\zeta)$ and $\varepsilon_1'(\lambda,\zeta)$ are exponentially small, since $\zeta_\infty(1+\varepsilon)$ is positive.

3. Uniform asymptotic formula for $p_n(x)e^{-v(x)/2}$. We first recall the asymptotic formula

$$\mathrm{Ai}(x) \sim \frac{1}{2\pi^{1/2}x^{1/4}} \exp\left(-\frac{2}{3}x^{3/2}\right)$$

and

$$\mathrm{Bi}(x) \sim \frac{1}{\pi^{1/2}x^{1/4}} \exp\left(\frac{2}{3}x^{3/2}\right)$$

as $x \to \infty$. Since the function $Y_n(x)$ in (1.11) is exponentially small as $x \to \infty$, by (2.5) and (2.58) there exists a constant $C(n)$ such that

$$Y_n(x) = C(n)\hat{H}_0(\lambda,w)^{-1/4} Z_1(\lambda,\zeta), \tag{3.1}$$

where Z_1 is the asymptotic solution given in (2.71). Substituting (1.11), (2.1), and (2.71) into (3.1), we have

$$\begin{aligned} p_n(x)e^{-v(x)/2} = \ & C(n)\lambda^{m-\frac{1}{2}}\left(\frac{\zeta}{w^2-1}\right)^{1/4}\left(1+\frac{\Phi'(\zeta)}{u}\right)^{-1/2} \\ & \cdot \{\mathrm{Ai}(X(\zeta)) + \varepsilon_1(\lambda,\zeta)\}, \end{aligned} \tag{3.2}$$

where $X(\zeta)$ is defined by

$$X(\zeta) = u^{2/3}\zeta + \frac{\Phi(\lambda,\zeta)}{u^{1/3}}. \tag{3.3}$$

Next we need to find a formula for $C(n)$ as $n \to \infty$. Put $x = 0$ and, equivalently, $w = 0$. Then (3.2) gives

$$\begin{aligned} p_n(0)e^{-v_0/2} = \ & C(n)\lambda^{m-\frac{1}{2}}(-\zeta(\lambda,0))^{1/4}\left(1+\frac{\Phi'(\lambda,\zeta(\lambda,0))}{u}\right)^{-1/2} \\ & \cdot \{\mathrm{Ai}[X(\zeta(\lambda,0))] + \varepsilon_1(\lambda,\zeta(\lambda,0))\}. \end{aligned} \tag{3.4}$$

First we consider the left-hand side of (3.4). From the recurrence relation (1.10), we have

$$p_n(0) = \begin{cases} 0, & n = 2k+1, \\ (-1)^k\gamma_0\dfrac{a_1a_3\cdots a_{2k-1}}{a_2a_4\cdots a_{2k}}, & n = 2k. \end{cases} \tag{3.5}$$

From (2.29), we also have

$$\begin{aligned} a_{n-1} &= a_n\left[1-\frac{1}{2mn}+O\left(\frac{1}{n^{1+1/m}}\right)\right] \\ &= a_n\left(1-\frac{1}{2mn}\right)\left[1+O\left(\frac{1}{n^{1+1/m}}\right)\right]. \end{aligned}$$

The last equation gives

$$\frac{a_{2k-1}}{a_{2k}} = \left(1-\frac{1}{4mk}\right)\left[1+O\left(\frac{1}{k^{1+1/m}}\right)\right].$$

Simple calculation shows

$$\prod_{k=1}^{n}\left(1-\frac{1}{4mk}\right)=\frac{\Gamma(n+1-\frac{1}{4m})}{\Gamma(n+1)\Gamma(1-\frac{1}{4m})} = n^{-1/4m}[1+O(n^{-1})]/\Gamma(1-\tfrac{1}{4m}). \tag{3.6}$$

Here use has been made of an asymptotic formula on the ratio of two gamma functions [7, p. 118]. It is easily seen that the infinite product $\prod_{k=1}^{\infty}[1+O(\frac{1}{k^{1+1/m}})]$ converges, say, to a finite value $\tilde{A}_1$. Then it is readily verifiable that

$$\prod_{k=1}^{n}\left[1+O\left(\frac{1}{k^{1+1/m}}\right)\right]=\tilde{A}_1\left[1+O\left(\frac{1}{n^{1/m}}\right)\right]. \tag{3.7}$$

Hence, when n is even, we obtain from (3.5)

$$p_n(0) = (-1)^{n/2}\gamma_0\prod_{k=1}^{n/2}\left(1-\frac{1}{4mk}\right)\prod_{k=1}^{n/2}\left[1+O\left(\frac{1}{k^{1+1/m}}\right)\right].$$

By (3.6) and (3.7), we get

$$p_n(0) = (-1)^{n/2}\tilde{A}_2 n^{-1/4m}[1+O(n^{-1/m})], \tag{3.8}$$

where $\tilde{A}_2 = 2^{1/4m}\gamma_0\tilde{A}_1/\Gamma(1-\frac{1}{4m})$.

Next we consider the right-hand side of (3.4). To simplify the notations, we put $\zeta_0 = \zeta(\lambda,0)$ and $X_0 = X(\zeta(\lambda,0))$. Note that when $\zeta \in [\zeta_\infty(-1+\varepsilon),\zeta_\infty(1-\varepsilon)] \subset (-\infty,0)$, or equivalently, $w \in [-1+\varepsilon,1-\varepsilon]$, we have $X(\zeta) = u^{2/3}\zeta+\Phi(\zeta)/u^{1/3} \to -\infty$ uniformly; thus, $X(\zeta_0) \to -\infty$. Recall the asymptotic formula [7, p. 392]

$$\begin{aligned} \mathrm{Ai}(-x) = \frac{1}{\pi^{1/2}x^{1/4}} \ &\left\{\cos\left(\frac{2}{3}x^{3/2}-\frac{\pi}{4}\right)[1+O(x^{-3})]\right. \\ &\left. +\ \sin\left(\frac{2}{3}x^{3/2}-\frac{\pi}{4}\right)\left[\frac{5}{48}x^{-3/2}+O(x^{-9/2})\right]\right\} \end{aligned} \tag{3.9}$$

as $x \to +\infty$. When ζ is bounded away from 0, $X(\zeta) = O(u^{2/3})$ and we thus have

$$\mathrm{Ai}(X(\zeta)) = \frac{1}{\pi^{1/2}[-X(\zeta)]^{1/4}}\left\{\cos\left(\frac{2}{3}[-X(\zeta)]^{3/2}-\frac{\pi}{4}\right)+O(u^{-1})\right\} \tag{3.10}$$

uniformly for $\zeta \in [\zeta_\infty(-1+\varepsilon),\zeta_\infty(1-\varepsilon)]$. In particular,

$$\mathrm{Ai}(X(\zeta_0)) = \frac{1}{\pi^{1/2}[-X(\zeta_0)]^{1/4}}\left\{\cos\left(\frac{2}{3}[-X(\zeta_0)]^{3/2}-\frac{\pi}{4}\right)+O(u^{-1})\right\}. \tag{3.11}$$

To proceed further, we need to derive an asymptotic formula for $\frac{2}{3}[-X(\zeta_0)]^{3/2}-\frac{\pi}{4}$, which is given in (3.25) below.

Recall from (2.56) that we have

$$\zeta(\lambda,w)=-\left[\frac{3}{2}\frac{1}{\lambda^{2m-1}}\int_w^1\sqrt{1-t^2}A_n(\lambda t)dt\right]^{2/3}$$

when $|w|<1$. Let

$$I(\lambda,w)=\frac{1}{\lambda^{2m-1}}\int_w^1\sqrt{1-t^2}A_n(\lambda t)dt=\frac{2}{3}[-\zeta(\lambda,w)]^{3/2}. \tag{3.12}$$

Inserting (2.33) in (3.12) yields

$$\begin{aligned}I(\lambda,w)=\frac{1}{\lambda^{2m-2}}&\sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k}\\&\cdot\sum_{l=0}^{k}\left\{2^{-2l}\binom{2l}{l}\int_w^1\sqrt{1-t^2}t^{2k-2l}dt\left[1+\frac{\tilde{d}_{2l}}{\lambda^{2m}}+O\left(\frac{1}{\lambda^{2m+2}}\right)\right]\right\}.\end{aligned} \tag{3.13}$$

Note that when $w=0$, the integral in (3.13) is a beta function and we have

$$\begin{aligned}\int_0^1\sqrt{1-t^2}\,t^{2k-2l}dt&=\frac{1}{2}B\left(k-l+\frac{1}{2},\frac{3}{2}\right)\\&=\frac{\Gamma(k-l+\frac{1}{2})\Gamma(\frac{3}{2})}{2\Gamma(k-l+2)}\\&=\frac{\pi}{2}\frac{1}{2^{2k-2l+1}(k-l+1)}\binom{2k-2l}{k-l}.\end{aligned}$$

Substituting this into (3.13), and then reindexing, we get

$$\begin{aligned}I(\lambda,0)=&\frac{\pi}{2}\frac{1}{\lambda^{2m}}\sum_{k=0}^{m-1}(k+1)v_{k+1}\frac{\lambda^{2k+2}}{2^{2k+1}}\sum_{l=0}^{k}\frac{1}{l+1}\binom{2k-2l}{k-l}\binom{2l}{l}\\&+\frac{\pi}{2}\frac{1}{\lambda^{4m}}\sum_{k=0}^{m-1}(k+1)v_{k+1}\frac{\lambda^{2k+2}}{2^{2k+1}}\sum_{l=0}^{k}\frac{\tilde{d}_{2l}}{l+1}\binom{2k-2l}{k-l}\binom{2l}{l}+O\left(\frac{1}{\lambda^{2m+2}}\right).\end{aligned} \tag{3.14}$$

For convenience, we put

$$D_{2k}=\sum_{l=0}^{k}\frac{\tilde{d}_{2l}}{l+1}\binom{2l}{l}\binom{2k-2l}{k-l}.$$

Using the combinatorial identity

$$\sum_{l=0}^{k}\frac{1}{l+1}\binom{2l}{l}\binom{2k-2l}{k-l}=\binom{2k+1}{k+1},$$

which can be proved by induction, (3.14) can be simplified to

$$\begin{aligned}I(\lambda,0)=&\frac{\pi}{2}\frac{1}{\lambda^{2m}}\sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k+2}2^{-(2k+1)}\binom{2k+1}{k+1}\\&+\frac{\pi}{2}\frac{1}{\lambda^{2m}}mv_m2^{-(2m-1)}D_{2m-2}+O\left(\frac{1}{\lambda^{2m+2}}\right).\end{aligned} \tag{3.15}$$

When $\zeta \in [\zeta_\infty(-1+\varepsilon), \zeta_\infty(1-\varepsilon)]$, $\Phi(\lambda,\zeta)$ is bounded for large λ. Hence it follows from (3.3) that

$$\begin{aligned}\frac{2}{3}[-X(\zeta)]^{3/2} &= \frac{2}{3}u(-\zeta)^{3/2}\left[1+\frac{3}{2}\frac{\Phi(\lambda,\zeta)}{\zeta u}+O(u^{-2})\right]\\ &= \lambda^{2m}I(\lambda,w)-(-\zeta)^{1/2}\Phi(\lambda,\zeta)+O(\lambda^{-2m}).\end{aligned}\tag{3.16}$$

When $w=0$, $\zeta=\zeta_0\neq 0$ and (3.16) becomes

$$\frac{2}{3}[-X(\zeta_0)]^{3/2}-\frac{\pi}{4}=\lambda^{2m}I(\lambda,0)-(-\zeta_0)^{1/2}\Phi(\lambda,\zeta_0)-\frac{\pi}{4}+O(\lambda^{-2m}).\tag{3.17}$$

Observe that the limit of the second term on the right-hand side of (3.17) exists as $\lambda\to\infty$. Inserting (3.15) in (3.17) gives

$$\begin{aligned}\frac{2}{3}[-X(\zeta_0)]^{3/2}-\frac{\pi}{4} &= \frac{\pi}{2}\sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k+2}2^{-(2k+1)}\binom{2k+1}{k+1}\\ &\qquad \frac{\pi}{2}mv_m2^{-(2m-1)}D_{2m-2}-(-\zeta_0)^{1/2}\Phi(\lambda,\zeta_0)-\frac{\pi}{4}+O(\lambda^{-2}).\end{aligned}\tag{3.18}$$

On the other hand, using integration by parts and orthogonality, we have

$$\begin{aligned}&\int_{-\infty}^{\infty}v'(x)p_n(x)p_{n-1}(x)e^{-v(x)}dx\\ &=\int_{-\infty}^{\infty}[p_n'(x)p_{n-1}(x)+p_n(x)p_{n-1}'(x)]e^{-v(x)}dx\\ &=n\frac{\gamma_n}{\gamma_{n-1}}\int_{-\infty}^{\infty}p_{n-1}^2(x)e^{-v(x)}dx=\frac{n}{a_n}.\end{aligned}\tag{3.19}$$

Substituting $v'(x)=2\sum_{k=1}^m kv_kx^{2k-1}$ into (3.19) and applying (2.9), we also have

$$\begin{aligned}&\int_{-\infty}^{\infty}v'(x)p_n(x)p_{n-1}(x)e^{-v(x)}dx\\ &=2\sum_{k=1}^{m}kv_kc_{n,1,2k-1}=2\sum_{k=0}^{m-1}(k+1)v_{k+1}c_{n,1,2k+1}.\end{aligned}\tag{3.20}$$

Coupling (3.19) and (3.20) gives

$$n=2a_n\sum_{k=0}^{m-1}(k+1)v_{k+1}c_{n,1,2k+1};\tag{3.21}$$

see [2, (1.2)]. Inserting (2.30) in (3.21), and noting that $2a_n=\lambda$, we have

$$n=\lambda\sum_{k=0}^{m-1}(k+1)v_{k+1}\binom{2k+1}{k+1}2^{-(2k+1)}\lambda^{2k+1}\left[1+\frac{\tilde c_{1,2k+1}}{2mn}+O\left(\frac{1}{n^{1+1/m}}\right)\right].$$

By (2.27) and the relationship between $\tilde c_{1,2k+1}$ and $\tilde d_{2k+1}$ given in the statement following (2.34), we obtain

$$\begin{aligned} n &= \sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k+2}2^{-(2k+1)}\binom{2k+1}{k+1}\left[1+\frac{\tilde{d}_{2k+1}}{\lambda^{2m}}+O\left(\frac{1}{\lambda^{2m+2}}\right)\right] \\ &= \sum_{k=0}^{m-1}(k+1)v_{k+1}\lambda^{2k+2}2^{-(2k+1)}\binom{2k+1}{k+1} \\ &\quad + mv_m 2^{-(2m-1)}\binom{2m-1}{m}\tilde{d}_{2m-1}+O(\lambda^{-2}). \end{aligned} \tag{3.22}$$

Recall from (2.25) that $L_m = \frac{m!(m-1)!}{(2m)!}$, and hence $L_k^{-1} = 2k\binom{2k-1}{k}$. Thus on one hand we have

$$\begin{aligned} n &= \sum_{k=1}^{m} v_k\lambda^{2k}2^{-2k}L_k^{-1} + 2^{-2m}L_m^{-1}\tilde{d}_{2m-1}+O(\lambda^{-2}) \\ &= 2^{-2m}L_m^{-1}u\left[1+v_{m-1}\frac{L_m}{L_{m-1}}2^{-2(m-1)}\frac{1}{u^{1/m}}+v_{m-2}\frac{L_m}{L_{m-2}}2^{-2(m-2)}\frac{1}{u^{2/m}}\right. \\ &\qquad \left. +\cdots+v_1\frac{L_m}{L_1}2^{-2}\frac{1}{u^{1-1/m}}+\frac{\tilde{d}_{2m-1}}{u}+O\left(\frac{1}{u^{1+1/m}}\right)\right], \end{aligned} \tag{3.23}$$

and on the other hand we have by (2.24)

$$u = 2^{2m}L_m n\left[1+\frac{\tau_1}{n^{1/m}}+\frac{\tau_2}{n^{2/m}}+\cdots+\frac{\tau_m}{n}+O\left(\frac{1}{n^{1+1/m}}\right)\right], \tag{3.24}$$

since $2a_n = \lambda$ and $\lambda^{2m} = u$. Inserting (3.24) in (3.23) and comparing the coefficients of $n^{-k/m}$ for $k = 0,\ldots,m$, we can determine the coefficients $\tau_1,\tau_2,\ldots,\tau_m$. The first two are given by

$$\begin{aligned} \tau_1 &= -2^{-2m}\frac{L_m}{L_{m-1}}L_m^{-1/m}v_{m-1}, \\ \tau_2 &= -2^{-2m}\frac{L_m}{L_{m-2}}L_m^{-2/m}v_{m-2}+\left(1-\frac{1}{m}\right)\tau_1^2. \end{aligned}$$

Since $a_n = u^{1/2m}/2$, we can also determine the coefficients $\eta_1,\eta_2,\ldots,\eta_m$ in (2.24). For instance, we have

$$\begin{aligned} \eta_1 &= -\frac{2^{-2m}}{2m}\frac{L_m}{L_{m-1}}L_m^{-1/m}v_{m-1}, \\ \eta_2 &= -\frac{2^{-2m}}{2m}\frac{L_m}{L_{m-2}}L_m^{-2/m}v_{m-2}+\left(m-\frac{3}{2}\right)\eta_1^2. \end{aligned}$$

Comparing (3.22) with (3.18) gives

$$\frac{2}{3}[-X(\zeta_0)]^{3/2}-\frac{\pi}{4} = \frac{\pi}{2}n+\beta(\lambda), \tag{3.25}$$

where

$$\begin{aligned} \beta(\lambda) &= \frac{\pi}{2}mv_m 2^{-(2m-1)}\left[D_{2m-2}-\binom{2m-1}{m}\tilde{d}_{2m-1}\right] \\ &\qquad -(-\zeta_0)^{1/2}\Phi(\lambda,\zeta_0)-\frac{\pi}{4}+O(\lambda^{-2}). \end{aligned} \tag{3.26}$$

Since $\zeta_0 = \zeta(\lambda, 0)$ and $\Phi(\lambda, \zeta_0)$ have limits as $\lambda \to \infty$, so does $\beta(\lambda)$ as $\lambda \to \infty$.

Substituting (3.3) and (3.25) in (3.10), we get

$$\text{(3.27)}\quad \mathrm{Ai}(X(\zeta_0)) = \pi^{-1/2}u^{-1/6}(-\zeta_0)^{-1/4}[1+O(u^{-1})]\left\{\cos\left(\frac{\pi}{2}n+\beta(\lambda)\right)+O(u^{-1})\right\},$$

which can be simplified to

$$\text{(3.28)}\qquad \mathrm{Ai}(X(\zeta_0)) = \pi^{-1/2}u^{-1/6}(-\zeta_0)^{-1/4}\left\{\cos\left(\frac{\pi}{2}n+\beta(\lambda)\right)+O(u^{-1})\right\}.$$

We need both of these results in our later discussions. Inserting (3.28) in (3.4) and noting that $\varepsilon_1(\lambda, \zeta_0) = O(u^{-7/6})$ and $u = \lambda^{2m}$, we have

$$\text{(3.29)}\qquad \begin{aligned} p_n(0)e^{-v_0/2} &= C(n)\pi^{-1/2}u^{1/3-1/4m} \\ &\quad \cdot \left\{\cos\left(\frac{\pi}{2}n\right)\cos\beta(\lambda) - \sin\left(\frac{\pi}{2}n\right)\sin\beta(\lambda)+O(u^{-1})\right\}. \end{aligned}$$

Now we differentiate both sides of (3.2) and then set $x = 0$ and, correspondingly, $w = 0$. This leads to

$$\text{(3.30)}\qquad \begin{aligned} &p_n'(0)\ \ e^{-v_0/2} = \\ &C(n)\lambda^{m-3/2}\left\{\frac{1}{4}(-\zeta_0)^{-3/4}(-\zeta_0')\left(1+\frac{\Phi'(\zeta_0)}{u}\right)^{-1/2}[\mathrm{Ai}(X(\zeta_0))+\varepsilon_1(\lambda,\zeta_0)]\right. \\ &-\frac{1}{2}(-\zeta_0)^{1/4}\left(1+\frac{\Phi'(\zeta_0)}{u}\right)^{-3/2}\frac{\Phi''(\zeta_0)}{u}\zeta_0'[\mathrm{Ai}(X(\zeta_0))+\varepsilon_1(\lambda,\zeta_0)] \\ &\left.+(-\zeta_0)^{1/4}\left(1+\frac{\Phi'(\zeta_0)}{u}\right)^{-1/2}\zeta_0'[\mathrm{Ai}'(X(\zeta_0))X'(\zeta_0)+\varepsilon_1'(\lambda,\zeta_0)]\right\}. \end{aligned}$$

By using the differential recurrence relation (1.5) and the fact that $B_n(x)$ is an odd function, we obtain

$$\text{(3.31)}\qquad p_n'(0) = A_n(0)p_{n-1}(0).$$

From (2.55) and (2.53), we also have

$$\text{(3.32)}\qquad \zeta_0' = \hat{H}_0(\lambda, 0)^{1/2} = (-\zeta_0)^{-1/2}\frac{A_n(0)}{\lambda^{2m-1}}.$$

Recall again that $u = \lambda^{2m}$. Hence, inserting (3.31) and (3.32) in (3.30) yields

$$\text{(3.33)}\qquad \begin{aligned} &p_{n-1}(0)e^{-v_0/2} = C(n)\lambda^{m-1/2}\left\{-\frac{1}{4u}(-\zeta_0)^{-5/4}\left(1+\frac{\Phi'(\zeta_0)}{u}\right)^{-1/2}[\mathrm{Ai}(X(\zeta_0))+\varepsilon_1(\lambda,\zeta_0)]\right. \\ &\qquad -\frac{1}{2}(-\zeta_0)^{-1/4}\left(1+\frac{\Phi'(\zeta_0)}{u}\right)^{-3/2}\frac{\Phi''(\zeta_0)}{u^2}[\mathrm{Ai}(X(\zeta_0))+\varepsilon_1(\lambda,\zeta_0)] \\ &\qquad \left.+\frac{1}{u}(-\zeta_0)^{-1/4}\left(1+\frac{\Phi'(\zeta_0)}{u}\right)^{-1/2}[\mathrm{Ai}'(X(\zeta_0))X'(\zeta_0)+\varepsilon_1'(\lambda,\zeta_0)]\right\}. \end{aligned}$$

The first term inside the curly brackets in (3.33) is of order $O(u^{-7/6})$ since $\mathrm{Ai}(X(\zeta_0)) = O(u^{-1/6})$ (cf. (3.28)), and the second term inside the brackets is of order $O(u^{-13/6})$. Since $X'(\zeta_0) = u^{2/3}(1 + \Phi'(\zeta_0)/u)$ by (3.3) and $\varepsilon_1'(\lambda, \zeta) = O(u^{-1/6})$ by (2.76), the third term is equal to

$$(-\zeta_0)^{-1/4}u^{-1/3}\left(1 + \frac{\Phi'(\zeta_0)}{u}\right)^{1/2}\mathrm{Ai}'(X(\zeta_0)) + O(u^{-7/6}).$$

Hence it follows from (3.33) that

$$\begin{aligned} p_{n-1}(0)e^{-v_0/2} &= C(n)\lambda^{m-1/2} \\ &\quad \cdot \left\{(-\zeta_0)^{-1/4}u^{-1/3}\left(1 + \frac{\Phi'(\zeta_0)}{u}\right)^{1/2}\mathrm{Ai}'(X(\zeta_0)) + O(u^{-7/6})\right\}. \end{aligned} \tag{3.34}$$

Using the asymptotic formula [7, p. 392]

$$\mathrm{Ai}'(-x) = \frac{x^{1/4}}{\pi^{1/2}}\left\{\sin\left(\frac{2}{3}x^{3/2} - \frac{\pi}{4}\right) + O(x^{-3/2})\right\}, \qquad x \to \infty,$$

we have from (3.3) and (3.25)

$$\mathrm{Ai}'(X(\zeta_0)) = \pi^{-1/2}(-\zeta_0)^{1/4}u^{1/6}\left\{\sin\left(\frac{\pi}{2}n + \beta(\lambda)\right) + O(u^{-1})\right\}; \tag{3.35}$$

cf. (3.28). Coupling (3.34) and (3.35) yields

$$\begin{aligned} p_{n-1}(0)e^{-v_0/2} &= C(n)\pi^{-1/2}u^{1/3-1/4m} \\ &\quad \cdot \left\{\sin\left(\frac{\pi}{2}n\right)\cos\beta(\lambda) + \cos\left(\frac{\pi}{2}n\right)\sin\beta(\lambda) + O(u^{-1})\right\}. \end{aligned} \tag{3.36}$$

When n is even, we have $p_{n-1}(0) = 0$, $\sin(\frac{\pi}{2}n) = 0$, and $\cos(\frac{\pi}{2}n) = (-1)^{n/2}$. Thus, from (3.36), we obtain $\sin\beta(\lambda) = O(u^{-1})$, from which it follows that $\cos\beta(\lambda) = \delta[1 + O(u^{-2})]$, where $\delta = 1$ or -1. When n is odd, we obtain $p_n(0) = 0$, $\cos(\frac{\pi}{2}n) = 0$, and $\sin(\frac{\pi}{2}n) = (-1)^{(n-1)/2}$. From (3.29), we again obtain $\sin\beta(\lambda) = O(u^{-1})$ and $\cos\beta(\lambda) = \delta[1 + O(u^{-2})]$. Since $\beta(\lambda)$ has a limit as $\lambda \to \infty$, we always have

$$\cos\beta(\lambda) = \delta[1 + O(u^{-2})] \qquad \text{as } \lambda \to \infty, \tag{3.37}$$

regardless of whether n is even or odd. Inserting (3.8) and (3.37) in (3.29) when n is even and in (3.36) when n is odd, we obtain

$$\tilde{A}_2 n^{-1/4m}e^{-v_0/2}[1 + O(n^{-1/m})] = C(n)\pi^{-1/2}\delta u^{1/3-1/4m}[1 + O(u^{-1})],$$

which in turn gives

$$C(n) = An^{-1/4m}u^{1/4m-1/3}[1 + O(u^{-1/m})], \tag{3.38}$$

where $A = \tilde{A}_2\delta^{-1}\pi^{1/2}e^{-v_0/2}$. Substituting (3.38) into (3.2) gives

$$p_n(x)e^{-v(x)/2} = An^{-1/4m}u^{1/6}\left(\frac{\zeta}{w^2 - 1}\right)^{1/4}\{\mathrm{Ai}(X(\zeta)) + \varepsilon_1(\lambda, \zeta)\}[1 + O(u^{-1/m})]. \tag{3.39}$$

This formula holds uniformly for $w \in [-1+\varepsilon, M]$.

Our next task is to determine the constant A. To this end, we restrict w to the interval $[-1+\varepsilon, 1-\varepsilon]$. By inserting (3.10), (2.76), and (3.3) in (3.39), we have

$$\begin{aligned} p_n(x)e^{-v(x)/2} = {} & A\pi^{-1/2}n^{-1/4m}(1-w^2)^{-1/4} \\ & \cdot \left\{\cos\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4}\right) + O(u^{-1})\right\}[1+O(u^{-1/m})] \end{aligned} \tag{3.40}$$

uniformly for $w \in [-1+\varepsilon, 1-\varepsilon]$, which can be simplified to

$$\begin{aligned} p_n(x)e^{-v(x)/2} = {} & A\pi^{-1/2}n^{-1/4m}(1-w^2)^{-1/4} \\ & \cdot \left\{\cos\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4}\right) + O(u^{-1/m})\right\}. \end{aligned} \tag{3.41}$$

Squaring both sides of the last equation gives

$$\begin{aligned} p_n^2(x)e^{-v(x)} = {} & A^2\pi^{-1}n^{-1/2m}(1-w^2)^{-1/2}\left\{\cos^2\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4}\right) + O(u^{-1/m})\right\} \\ = {} & \frac{1}{2}A^2\pi^{-1}n^{-1/2m}(1-w^2)^{-1/2}\left\{\sin\left(\frac{4}{3}[-X(\zeta)]^{3/2}\right) + 1 + O(u^{-1/m})\right\}. \end{aligned} \tag{3.42}$$

On account of (1.2) and (2.4), we have

$$1 = \int_{-\infty}^{\infty} p_n^2(x)e^{-v(x)}dx \geq \lambda\int_{-1+\varepsilon}^{1-\varepsilon} p_n^2(x)e^{-v(x)}dw.$$

From (3.42), it follows that

$$\begin{aligned} 1 \geq \frac{\lambda}{2}A^2\pi^{-1}n^{-1/2m} & \left[\int_{-1+\varepsilon}^{1-\varepsilon} \frac{1}{\sqrt{1-w^2}}\sin\left(\frac{4}{3}[-X(\zeta)]^{3/2}\right)dw\right. \\ & \left. + \int_{-1+\varepsilon}^{1-\varepsilon} \frac{dw}{\sqrt{1-w^2}} + O(u^{-1/m})\right]. \end{aligned} \tag{3.43}$$

Let us consider for a moment the function

$$\xi = \xi(w) = \frac{4}{3}(-\zeta)^{3/2}\left(1 + \frac{\Phi(\zeta)}{u\zeta}\right)^{3/2}.$$

Clearly

$$\begin{aligned} \xi'(w) = {} & -2(-\zeta)^{1/2}\left(1 + \frac{\Phi(\zeta)}{u\zeta}\right)^{3/2}\zeta'(\lambda, w) \\ & + 2(-\zeta)^{3/2}\left(1 + \frac{\Phi(\zeta)}{u\zeta}\right)^{1/2}\frac{1}{u}\frac{d}{dw}\frac{\Phi(\zeta)}{\zeta}. \end{aligned}$$

The first term on the right-hand side is negative for $w \in [-1+\varepsilon, 1-\varepsilon]$ and for λ sufficiently large, the second term tends to zero, uniformly for $w \in [-1+\varepsilon, 1-\varepsilon]$, as $\lambda \to \infty$. Thus, when λ is sufficiently large, we have $\xi'(w) < 0$. Therefore, the mapping $w \mapsto \xi$ is one-to-one on the interval $[-1+\varepsilon, 1-\varepsilon]$. Since

$$\frac{4}{3}[-X(\zeta)]^{3/2} = u \cdot \frac{4}{3}(-\zeta)^{3/2}\left(1+\frac{\Phi(\zeta)}{u\zeta}\right)^{3/2},$$

the first integral inside the square brackets in (3.43) is equal to

$$I_0 = \int_{\xi(-1+\varepsilon)}^{\xi(1-\varepsilon)} \frac{1}{\sqrt{1-\omega^2(\xi)}}\frac{1}{\xi'(\omega(\xi))}\sin(u\xi)d\xi,$$

where $\omega(\xi)$ is the inverse of $\xi = \xi(w)$. The Riemann–Lebesgue lemma infers that this integral tends to 0 as $\lambda \to \infty$, or equivalently, as $u \to \infty$. By (2.27),

$$\lambda n^{-1/2m} \to 2L_m^{1/2m} \tag{3.44}$$

as $\lambda \to \infty$. Hence, letting $\lambda \to \infty$ in (3.43), we obtain

$$1 \geq 2A^2\pi^{-1}L_m^{1/2m} \cdot \arcsin(1-\varepsilon).$$

Since ε is arbitrary, it follows that

$$A^2 \leq L_m^{-1/2m}. \tag{3.45}$$

To show that the reverse inequality also holds, we consider the identity

$$\int_{-\infty}^{\infty}\left(1-\frac{x^2}{\lambda^2}\right)p_n^2(x)e^{-v(x)}dx = 1-\frac{1}{\lambda^2}(a_{n+1}^2+a_n^2), \tag{3.46}$$

which can be obtained by squaring both sides of the recurrence relation (1.10) and using the orthonormal property (1.2). From (2.4) and (2.54), it is evident that

$$\int_{-\infty}^{\infty}\left(1-\frac{x^2}{\lambda^2}\right)p_n^2(x)e^{-v(x)}dx \leq 2\lambda\int_0^1(1-w^2)p_n^2(x)e^{-v(x)}dw.$$

Divide the interval of integration on the right-hand side at $w = 1-\varepsilon n^{-\alpha}$, $\alpha > 0$, and denote by I_1 and I_2 the integrals corresponding, respectively, to the subintervals $[0, 1-\varepsilon n^{-\alpha}]$ and $[1-\varepsilon n^{-\alpha}, 1]$. Thus,

$$\int_{-\infty}^{\infty}\left(1-\frac{x^2}{\lambda^2}\right)p_n^2(x)e^{-v(x)}dx \leq 2\lambda I_1 + 2\lambda I_2. \tag{3.47}$$

Since $\zeta(\lambda, w)$ is continuous in $[-1+\varepsilon, M]$ and has limit as $\lambda \to \infty$, $\zeta(\lambda, w)$ is uniformly bounded there. By the same reasoning, $\zeta/(w^2-1)$ is uniformly bounded in this interval. Furthermore, in view of (3.3) and (2.75), $\mathrm{Ai}(X(\zeta))$ and $\varepsilon_1(\lambda,\zeta)$ are also uniformly bounded in $[-1+\varepsilon, M]$. Hence, by virtue of (3.39), there exists a constant K such that

$$p_n^2(x)e^{-v(x)} \leq A^2n^{-1/2m}u^{1/3}K$$

uniformly for $w \in [-1+\varepsilon, M]$, from which it follows that

$$\begin{aligned} 0 \leq 2\lambda I_2 &= 2\lambda\int_{1-\varepsilon u^{-\alpha}}^{1}(1-w^2)p_n^2(x)e^{-v(x)}dw \\ &\leq 2\lambda A^2n^{-1/2m}u^{1/3}K\varepsilon u^{-\alpha}(2\varepsilon u^{-\alpha}-\varepsilon^2u^{-2\alpha}) \\ &= 8A^2KL_m^{1/2m}\varepsilon^2u^{1/3-2\alpha}[1+O(1)], \end{aligned} \tag{3.48}$$

where we have also made use of (3.44). We choose $\alpha > 1/6$ so that $2\lambda I_2 \to 0$ as $\lambda \to \infty$.

Now consider the integral I_1. When $w \in (0,1)$, by the mean-value theorem, $\zeta(\lambda, w) = \zeta'(\lambda, \omega_\lambda)(w-1)$, where $\omega_\lambda \in (w,1)$. Since $\zeta'(\lambda, w) \to \zeta'_\infty(w)$ uniformly on [0,1] and $\zeta'_\infty(w) > 0$ on [0,1] (see section 2), there exists a constant $K_1 > 0$ such that $-\zeta(\lambda, w) > K_1(1-w)$ uniformly for $w \in (0,1)$ and for all sufficiently large λ. Hence, when $0 < w < 1 - \varepsilon u^{-\alpha}$, we have $-\zeta(\lambda, w) \geq K_1 \varepsilon u^{-\alpha}$, from which it follows from (3.3) that

$$-X(\zeta) \geq K_1 \varepsilon u^{2/3-\alpha}(1 - |\Phi(\zeta)| u^{\alpha-1}/K_1\varepsilon) \geq \frac{1}{2} K_1 \varepsilon u^{2/3-\alpha}$$

for sufficiently large u and for $0 < \alpha < 1$ since $\Phi(\zeta)$ is bounded. Hence, if $\alpha < 2/3$, then $-X(\zeta) \to +\infty$. Also, we have $-X(\zeta) = u^{2/3}(-\zeta)[1 + O(u^{\alpha-1})]$ and $[-X(\zeta)]^{-3/2} = O(u^{3\alpha/2-1})$ uniformly for $w \in [0, 1-\varepsilon u^{-\alpha}]$. By (3.9), we have for $1/6 < \alpha < 2/3$

$$\begin{aligned} \mathrm{Ai}(X(\zeta)) &= \pi^{-1/2} u^{-1/6} (-\zeta)^{-1/4} [1 + O(u^{\alpha-1})] \\ &\quad \cdot \left\{ \cos\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4} \right) + O(u^{3\alpha/2-1}) \right\}, \end{aligned} \tag{3.49}$$

which may be simplified to

$$\mathrm{Ai}(X(\zeta)) = \pi^{-1/2} u^{-1/6} (-\zeta)^{-1/4} \left\{ \cos\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4} \right) + O(u^{-1/2}) \right\} \tag{3.50}$$

with $\alpha = 1/3$. Substituting (3.50) into (3.39) and squaring both sides give

$$p_n^2(x) e^{-v(x)} = A^2 \pi^{-1} n^{-1/2m} (1-w^2)^{-1/2} \left\{ \cos^2\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4} \right) + O(u^{-1/m}) \right\}. \tag{3.51}$$

Inserting this in I_1, we obtain

$$2\lambda I_1 = 2\lambda \pi^{-1} A^2 n^{-1/2m} \int_0^{1-\varepsilon u^{-\alpha}} \sqrt{1-w^2} \left\{ \cos^2\left(\frac{2}{3}[-X(\zeta)]^{3/2} - \frac{\pi}{4} \right) + O(u^{-1/m}) \right\} dw,$$

which in turn gives

$$\begin{aligned} 2\lambda I_1 = \lambda \pi^{-1} A^2 n^{-1/2m} &\left[\int_0^{1-\varepsilon u^{-\alpha}} \sqrt{1-w^2} \sin\left(\frac{4}{3}[-X(\zeta)]^{3/2} \right) dw \right. \\ &\left. + \int_0^{1-\varepsilon u^{-\alpha}} \sqrt{1-w^2} dw + O(u^{-1/m}) \right]. \end{aligned} \tag{3.52}$$

As before, the Riemann–Lebesgue lemma infers that the first integral inside the square brackets in (3.52) tends to 0 as $\lambda \to \infty$. From (3.44) and the result

$$\int \sqrt{1-w^2} dw = \frac{1}{2} \left(w\sqrt{1-w^2} + \arcsin w \right),$$

it follows that

$$2\lambda I_1 \to \frac{1}{2} A^2 L_m^{1/2m} \qquad \text{as } \lambda \to \infty.$$

Since $\lambda = 2a_n$ and $a_{n+1} \sim a_n$, we have $1 - \frac{1}{\lambda^2}(a_{n+1}^2 + a_n^2) \to \frac{1}{2}$ as $\lambda \to \infty$. In view of (3.46) and (3.47), we get

$$1 \le A^2 L_m^{1/2m}. \tag{3.53}$$

Coupling (3.45) and (3.53) yields

$$A = L_m^{-1/4m}. \tag{3.54}$$

From (3.39), we obtain

$$\begin{aligned} p_n(x)e^{-v(x)/2} &= L_m^{-1/4m} n^{-1/4m} u^{1/6} \\ &\quad \cdot \left(\frac{\zeta}{w^2-1}\right)^{1/4} \{\mathrm{Ai}(X(\zeta)) + \varepsilon_1(\lambda,\zeta)\}[1 + O(u^{-1/m})]. \end{aligned} \tag{3.55}$$

Since $\varepsilon_1(\lambda,\zeta) = O(u^{-1})$ uniformly on $[-1+\varepsilon, M]$ by (2.75) and $u = 2^{2m}L_m n[1 + O(n^{-1/m})]$ by (2.24), we have from (3.55) the following main result.

THEOREM 3.1. *For any positive and fixed numbers $\varepsilon \ll 1$ and $1 \ll M$, the asymptotic formula*

$$p_n(x)e^{-v(x)/2} = 2^{m/3} L_m^{1/6-1/4m} n^{1/6-1/4m} \left(\frac{\zeta}{w^2-1}\right)^{1/4} \{\mathrm{Ai}(X(\zeta)) + O(n^{-1/m})\} \tag{3.56}$$

holds uniformly for $w \in [-1+\varepsilon, M]$, where $x = \lambda w$, ζ is defined in (2.56), *and $X(\zeta)$ is defined in* (3.3).

4. Asymptotic formula for the zeros of $p_n(x)$.

Let the zeros of $p_n(x)$ be arranged in decreasing order:

$$-\infty < x_{n,n} < x_{n,n-1} < \cdots < x_{n,2} < x_{n,1} < \infty, \tag{4.1}$$

and let $w_{n,k}$ and $\zeta_{n,k}$ denote the corresponding values determined by (2.4) and (2.56), respectively. Since $\zeta(\lambda, w)$ is an increasing function for $w \in (-1, \infty)$, (4.1) implies

$$-\infty < \zeta_{n,n} < \zeta_{n,n-1} < \cdots < \zeta_{n,2} < \zeta_{n,1} < \infty. \tag{4.2}$$

In view of the fact that

$$\lim_{w \to 1} \frac{\zeta}{w^2-1} = \frac{1}{2}\zeta'(\lambda, 1) > 0,$$

it follows from (3.55) that $\zeta_{n,k}$ is the kth root of the equation

$$\mathrm{Ai}(X(\zeta)) + \varepsilon_1(\lambda,\zeta) = 0, \tag{4.3}$$

where $X(\zeta)$ is defined in (3.3) and $\varepsilon_1(\lambda,\zeta)$ satisfies the estimates in (2.73). Let $\tilde{a}_k$ denote the kth negative zero of the Airy function $\mathrm{Ai}(x)$. Since $\varepsilon_1(\lambda,\zeta) = O(u^{-1})$, it is reasonable to expect that

$$\zeta_{n,k} \approx \zeta_n(\tilde{a}_k),$$

where $\zeta_n(\tilde{a}_k)$ is uniquely determined by

$$X(\zeta_n(\tilde{a}_k)) = u^{2/3}\zeta_n(\tilde{a}_k) + \frac{\Phi(\lambda, \zeta_n(\tilde{a}_k))}{u^{1/3}} = \tilde{a}_k \tag{4.4}$$

since $X(\zeta)$ is monotonically increasing for large λ. Note that $\zeta = \zeta(\lambda, w)$ is uniformly bounded on $[-1+\varepsilon, M]$ and $\Phi(\lambda,\zeta)$ is also uniformly bounded for $\zeta \in [\zeta_\infty(-1+\varepsilon), \zeta_\infty(M)]$. Hence, it follows from (4.4) that $\zeta_n(\tilde{a}_k)$ is negative when λ (or, equivalently, n) is sufficiently large.

Now we consider equation (4.3) and recall the estimate in (2.73). In what follows, we shall suppose that λ is sufficiently large so that

$$\frac{K\pi}{u}\mathcal{V}_{\zeta,\zeta_\infty(M)}(\Psi(\xi)) \exp\left\{\frac{K_0}{u}\mathcal{V}_{\zeta,\zeta_\infty(M)}(\Psi(\xi))\right\} < \frac{1}{2}. \tag{4.5}$$

For convenience, we put

$$\rho_1(\lambda,\zeta) = \varepsilon_1(\lambda,\zeta)E(X(\zeta))/M(X(\zeta)) \tag{4.6}$$

and

$$\sigma_1(\lambda,\zeta) = \frac{K\pi}{u}\mathcal{V}_{\zeta,\zeta_\infty(M)}(\Psi(\xi)) \exp\left\{\frac{K_0}{u}\mathcal{V}_{\zeta,\zeta_\infty(M)}(\Psi(\xi))\right\}. \tag{4.7}$$

From (2.73) and (4.5), we have $|\rho_1(\lambda,\zeta)| \le \sigma_1(\lambda,\zeta) < 1/2$. In terms of the phase function $\theta(x)$ associated with $\mathrm{Ai}(x)$, (4.3) can be written as

$$\sin\theta(X(\zeta)) = -\rho_1(\lambda,\zeta) \tag{4.8}$$

on account of (4.6). From the definition of the phase function, it is readily seen that $\theta(x) = \pi/4$ when $x \ge c$. Hence, for $X(\zeta) \ge c$, we have $\sin\theta(X(\zeta)) = 1/\sqrt{2}$. However, the absolute value of the right-hand side of (4.8) is less than 1/2. Thus, it follows that (4.3) has no roots when $X(\zeta) \ge c$. That is, all the roots $\zeta_{n,k}$ of (4.3) lie in the interval $X(\zeta) < c$. Since $X(\zeta) = u^{2/3}\zeta + \Phi(\zeta)/u^{1/3} \ge c$ when $\zeta \ge 0$ and λ is sufficiently large, (4.3) has no roots for $\zeta \ge 0$ and λ sufficiently large. Recall that $\zeta \ge 0$ corresponds to $w \ge 1$ by (2.56), and that $w \ge 1$ corresponds to $x \ge \lambda$ by (2.4). Therefore, in view of (3.55), the polynomial $p_n(x)$ has no zero in $x \ge \lambda$. By symmetry, $p_n(x)$ also has no zero in $x \le -\lambda$. That is, all zeros of $p_n(x)$ lie in the interval $-\lambda < x < \lambda$; equivalently, the corresponding values $\zeta_{n,k}$ all lie in $-\zeta(\lambda,-1) < \zeta < 0$. In this range and when $X(\zeta) < c$, (4.8), which is equivalent to (4.3), can be written as

$$\theta(X(\zeta)) - k\pi - (-1)^{k-1}\arcsin\{\rho_1(\lambda,\zeta)\} = 0, \tag{4.9}$$

where k is an arbitrary integer. In what follows, we shall show that the left-hand side of (4.9) is a decreasing function of ζ for $X(\zeta) < c$. In view of the identity [7, p. 404]

$$\theta'(x) = -1/\{\pi M^2(x)\} \qquad \text{for} \quad x \le c \tag{4.10}$$

and the fact that $M(x) \ne 0$, $\theta'(x)$ is strictly negative for all $x \le c$. Thus, to prove that the derivative of the function on the left-hand side of (4.9) is strictly less than zero, it suffices to show that

$$\{1 - \rho_1^2(\lambda,\zeta)\}^{-1/2}|\rho_1'(\lambda,\zeta)| < |\theta'(X(\zeta))|X'(\zeta). \tag{4.11}$$

Also, since $E(x) = 1$ for $x \leq c$, by (4.6) we have

$$\rho_1'(\lambda, \zeta) = \frac{\varepsilon_1'(\lambda, \zeta)}{M(X(\zeta))} - \frac{M'(X(\zeta))}{M^2(X(\zeta))} X'(\zeta)\varepsilon_1(\lambda, \zeta).$$

From (2.73) and (4.7), it follows that

$$\begin{aligned} |\rho_1'(\lambda, \zeta)| \leq & \frac{N(X(\zeta))X'(\zeta)\sigma_1(\lambda, \zeta)}{M(X(\zeta))} + \frac{M'(X(\zeta))X'(\zeta)}{M(X(\zeta))}\sigma_1(\lambda, \zeta) \\ = & \{N(X(\zeta)) + M'(X(\zeta))\} \frac{X'(\zeta)\sigma_1(\lambda, \zeta)}{M(X(\zeta))} \end{aligned}$$

and

$$\{1 - \rho_1^2(\lambda, \zeta)\}^{-1/2} |\rho_1'(\lambda, \zeta)| \leq \frac{\{N(X(\zeta)) + M'(X(\zeta))\}}{\{1 - \rho_1^2(\lambda, \zeta)\}^{1/2}} \frac{X'(\zeta)\sigma_1(\lambda.\zeta)}{M(X(\zeta))}. \tag{4.12}$$

Hence, it is evident that (4.11) holds if

$$\frac{\{N(X(\zeta)) + M'(X(\zeta))\}}{\{1 - \rho_1^2(\lambda, \zeta)\}^{1/2}} \frac{X'(\zeta)\sigma_1(\lambda.\zeta)}{M(X(\zeta))} < |\theta'(X(\zeta))| X'(\zeta).$$

By (4.10) and the estimate $|\rho_1(\lambda, \zeta)| \leq \sigma_1(\lambda, \zeta)$, we only need to prove that

$$\frac{\sigma_1(\lambda, \zeta)}{\{1 - \sigma_1^2(\lambda, \zeta)\}^{1/2}} < \frac{1}{\pi M(X(\zeta))\{N(X(\zeta)) + M'(X(\zeta))\}}. \tag{4.13}$$

The left-hand side of (4.13) is less than $1/\sqrt{3} = 0.577\ldots$, since $\sigma_1(\lambda, \zeta) < 1/2$. The right-hand side of (4.13) is a decreasing function of $X(\zeta)$ by Lemma 5.1 in [7, p. 404]. When $X(\zeta) = c$, its value is

$$\frac{1}{\pi \operatorname{Ai}(c) \left\{ \operatorname{Ai}'(c) + \operatorname{Bi}'(c) + \sqrt{2 \operatorname{Ai}'^2(c) + 2 \operatorname{Bi}'^2(c)} \right\}} = 0.708\ldots.$$

Therefore, (4.13) and (4.11) are both satisfied. This proves that the left-hand side of (4.9) is monotonically decreasing in ζ for every k.

From the estimate

$$|\arcsin \rho_1(\lambda, \zeta)| < \arcsin \frac{1}{2} = \frac{\pi}{6} \tag{4.14}$$

and the fact that

$$\theta(c) = \frac{\pi}{4},$$

one easily verifies that if $k \leq 0$, then the value of the left-hand side of (4.9) is greater than zero when $X(\zeta) = c$. Since the left-hand side of (4.9) is a decreasing function, (4.9) has no roots when $X(\zeta) \leq c$.

Let b_k denote the kth negative zero of $\operatorname{Bi}(x)$, and let $\zeta_n(b_k)$ be uniquely determined by

$$X(\zeta_n(b_k)) = b_k. \tag{4.15}$$

For any given positive integer k, from (4.14) and the result [7, p. 404]

$$\theta(b_k) = (k - \frac{1}{2})\pi,$$

it is readily seen that the left-hand side of (4.9) is negative when $\zeta = \zeta_n(b_k)$ and positive when $\zeta = \zeta_n(b_{k+1})$. Hence, in the range

$$b_{k+1} < X(\zeta) < b_k \tag{4.16}$$

or, equivalently, $\zeta_n(b_{k+1}) < \zeta < \zeta_n(b_k)$, (4.9) has a root $\tilde{\zeta}_{n,k}$. Since the left-hand side of (4.9) is decreasing in ζ, this is the only root of (4.9) for a fixed k, and the sequence $\tilde{\zeta}_{n,n} < \tilde{\zeta}_{n,n-1} < \cdots < \tilde{\zeta}_{n,1}$ includes all the roots of (4.8). Comparing with (4.2) and (4.3), we conclude that $\tilde{\zeta}_{n,k}$ must be equal to $\zeta_{n,k}$.

Next we investigate the relationship between the zero $\zeta_{n,k}$ and the kth zero $\tilde{a}_k$ of $\mathrm{Ai}(x)$. For $\zeta = \zeta_{n,k}$, we have, by the mean-value theorem,

$$\theta(X(\zeta)) = \theta(\tilde{a}_k) + (X(\zeta) - \tilde{a}_k)\theta'(\xi),$$

where $\xi \in (b_{k+1}, b_k)$. Recall from [7, p. 404] that $\theta(\tilde{a}_k) = k\pi$. Hence, by (4.9),

$$X(\zeta) - \tilde{a}_k = (-1)^{k-1} \arcsin\{\rho_1(\lambda, \zeta)\}/\theta'(\xi). \tag{4.17}$$

Using the inequalities $|\rho_1(\lambda, \zeta)| \leq \sigma_1(\lambda, \zeta) < 1/2$ and $\sin t > 3t/\pi$ for $0 < t < \pi/6$, we obtain

$$|X(\zeta) - \tilde{a}_k| \leq \frac{\pi}{3}\sigma_1(\lambda, \zeta)/|\theta'(\xi)|. \tag{4.18}$$

Since $|\theta'(\xi)|$ is decreasing in ξ (see [7, p. 404]) and $\sigma_1(\lambda, \zeta)$ is decreasing in ζ, it follows that

$$|X(\zeta) - \tilde{a}_k| \leq \alpha_k, \tag{4.19}$$

where

$$\alpha_k = \frac{\pi}{3}\sigma_1(\lambda, \zeta_n(b_{k+1}))/|\theta'(b_k)|. \tag{4.20}$$

(Recall that here $\zeta = \zeta_{n,k} \in (\zeta_n(b_{k+1}), \zeta_n(b_k))$.) In view of (4.10) and (4.7), (4.20) gives

$$\alpha_k = \frac{\pi^2}{3}M^2(b_k)\sigma_1(\lambda, \zeta_n(b_{k+1})) = M^2(b_k)O(u^{-1}). \tag{4.21}$$

Much of the above argument is patterned after that given in [7, pp. 406–407]. Coupling (4.21) with (4.19) and noting that $X(\zeta) = u^{2/3}\zeta + \Phi(\zeta)/u^{1/3}$ and $\zeta = \zeta_{n,k}$, we obtain

$$u^{2/3}\zeta_{n,k} + \frac{\Phi(\zeta_{n,k})}{u^{1/3}} = \tilde{a}_k + O(u^{-1}). \tag{4.22}$$

Since $\Phi(\zeta_{n,k})$ is bounded for all n and k, we have the preliminary approximation $\zeta_{n,k} = \tilde{a}_k u^{-2/3} + O(u^{-1})$. Note that for any fixed k, $\zeta_{n,k} \to 0$ as $n \to \infty$ (or, equivalently, as $u \to \infty$). By the mean-value theorem,

$$\Phi(\zeta_{n,k}) = \Phi(0) + \Phi'(\xi)\zeta_{n,k} = \Phi(0) + O(u^{-2/3}), \tag{4.23}$$

where $\xi_{n,k} < \xi < 0$ and we have used the boundedness of $\Phi'(\xi)$. Substituting (4.23) into (4.22) gives

$$\zeta_{n,k} = \tilde{a}_k u^{-2/3} - \Phi(0)u^{-1} + O(u^{-5/3}). \tag{4.24}$$

Recall that $\Phi(\zeta) = \Phi(\lambda, \zeta)$ depends on λ (or $u = \lambda^{2m}$); see (2.68). Hence we need to find an approximate value for $\Phi(0)$. From the equation following (2.68), we have

$$\Phi(0) = \phi(\lambda, 0), \tag{4.25}$$

where $\phi(\lambda, \zeta)$ is given in (2.62). In view of (2.53),

$$\phi(\lambda, \zeta) = \frac{H_1(\lambda, w)}{H_0(\lambda, w)}\zeta = \frac{\lambda^{4m-2}H_1(\lambda, w)}{(w+1)A_n^2(\lambda w)}\frac{\zeta}{w-1}. \tag{4.26}$$

Since $\zeta = 0$ corresponds to $w = 1$, putting $\zeta = 0$ and $w = 1$ in (4.26) yields

$$\phi(\lambda, 0) = \frac{\lambda^{4m-2}H_1(\lambda, 1)}{2A_n^2(\lambda)}\zeta'(\lambda, 1). \tag{4.27}$$

In accordance with (2.47) and the definition $H_1(\lambda, w) = \lambda^{-2m}U_1(\lambda w, n)$,

$$H_1(\lambda, 1) = \lambda^{-2m+1}A_n(\lambda)\left\{\lambda\left[A_n(\lambda) - A_{n-1}(\lambda)\frac{a_n}{a_{n-1}} + B_n(\lambda) - B_{n+1}(\lambda)\right] - 1\right\}. \tag{4.28}$$

Coupling (2.33) and the equation following (2.37) yields

$$\begin{aligned}&\lambda\left[A_n(\lambda) - A_{n-1}(\lambda)\frac{a_n}{a_{n-1}}\right]\\ &= \frac{2^{2m}L_m}{m\lambda^{2m-1}}\sum_{k=1}^{m-1}(k+1)v_{k+1}\lambda^{2k+1}\sum_{l=1}^{k}2^{-2l}l\binom{2l}{l} + O(\lambda^{-2}).\end{aligned} \tag{4.29}$$

In a similar manner, one can show by using (2.34) that

$$\begin{aligned}&\lambda\left[B_n(\lambda) - B_{n+1}(\lambda)\right]\\ &= -\frac{2^{2m}L_m}{m\lambda^{2m-1}}\sum_{k=1}^{m-1}(k+1)v_{k+1}\lambda^{2k+1}\sum_{l=0}^{k-1}2^{-2l-1}(l+1)\binom{2l+1}{l+1} + O(\lambda^{-2}).\end{aligned} \tag{4.30}$$

Since the leading terms on the right-hand side of (4.29) and (4.30) are equal, we have

$$\lambda\left[A_n(\lambda) - A_{n-1}(\lambda)\frac{a_n}{a_{n-1}} + B_n(\lambda) - B_{n+1}(\lambda)\right] = O(\lambda^{-2}). \tag{4.31}$$

Inserting (4.31) in (4.28) gives

$$\frac{\lambda^{2m-1}H_1(\lambda, 1)}{A_n(\lambda)} = -1 + O(\lambda^{-2}). \tag{4.32}$$

A combination of (4.25), (4.27), (4.32), and the equation preceding (2.68) leads to

$$\Phi(0) = -2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} + O(\lambda^{-2}). \tag{4.33}$$

Coupling (4.33) and (4.24) gives

$$\zeta_{n,k} = \tilde{a}_k u^{-2/3} + 2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} u^{-1} + O\left(u^{-(1+1/m)}\right). \tag{4.34}$$

Let $w = \Omega(\zeta) = \Omega(\lambda, \zeta)$ denote the inverse of the function $\zeta = \zeta(w) = \zeta(\lambda, w)$. Note that $\zeta_{n,k} = \zeta(w_{n,k})$ and $x_{n,k} = \lambda w_{n,k} = u^{1/2m} w_{n,k}$. Hence, it follows from (4.34) that

$$x_{n,k} = u^{1/2m}\Omega\left\{\tilde{a}_k u^{-2/3} + 2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} u^{-1} + O\left(u^{-(1+1/m)}\right)\right\}.$$

By the mean-value theorem, we obtain

$$x_{n,k} = u^{1/2m}\Omega\left\{\tilde{a}_k u^{-2/3} + 2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} u^{-1}\right\} + O\left(u^{-(1+1/2m)}\right). \tag{4.35}$$

To proceed further, we expand $\Omega(\zeta)$ into the Maclaurin series

$$\Omega(\zeta) = \Omega(0) + \Omega'(0)\zeta + \frac{1}{2}\Omega''(0)\zeta^2 + \cdots. \tag{4.36}$$

Since $\zeta(1) = 0$, we have $\Omega(0) = 1$. By the equation preceding (2.68),

$$\Omega'(0) = \frac{1}{\zeta'(1)} = 2^{-1/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-2/3}. \tag{4.37}$$

To get the value of $\Omega''(0)$, we need to find $\zeta''(1)$. By continuity, $\zeta^{(k)}(1) = \lim_{w\to 1}\zeta^{(k)}(w)$ for any positive integer k. Let $f(\lambda, w)$ and $g(\lambda, w)$ be given as in (2.66) and (2.67), respectively. By differentiating both sides of the equation

$$\zeta(w) = (w-1)\left[\frac{3}{2}g(\lambda, w)\right]^{2/3} \tag{4.38}$$

twice, we obtain

$$\zeta''(1) = 2\left[\frac{3}{2}g(\lambda, 1)\right]^{-1/3} g'(\lambda, 1). \tag{4.39}$$

From an equation in the proof of Lemma 2.2, we also have

$$g(\lambda, 1) = \frac{2}{3}f(\lambda, 1) = \frac{2\sqrt{2}}{3\lambda^{2m-1}}A_n(\lambda) \tag{4.40}$$

and

$$g'(\lambda, 1) = \frac{2}{5}f'(\lambda, 1) = \frac{1}{5\sqrt{2}\lambda^{2m-1}}A_n(\lambda) + \frac{2\sqrt{2}}{5\lambda^{2m-1}}\lambda A_n'(\lambda); \tag{4.41}$$

cf. (2.66). A combination of (4.40), (4.41), and (4.39) gives

$$\zeta''(1) = \frac{\sqrt{2}}{5}\left[\frac{A_n(\lambda)}{\lambda^{2m-1}} + \frac{4A_n'(\lambda)}{\lambda^{2m-2}}\right]\left[\frac{\sqrt{2}A_n(\lambda)}{\lambda^{2m-1}}\right]^{-1/3}, \tag{4.42}$$

and hence

$$\Omega''(0) = -\frac{\zeta''(1)}{\zeta'(1)^3} = -\frac{\sqrt{2}}{5}\left[\frac{A_n(\lambda)}{\lambda^{2m-1}} + \frac{4A_n'(\lambda)}{\lambda^{2m-2}}\right]\left[\frac{\sqrt{2}A_n(\lambda)}{\lambda^{2m-1}}\right]^{-7/3}. \tag{4.43}$$

Taking

$$\zeta = \tilde{a}_k u^{-2/3} + 2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} u^{-1} \tag{4.44}$$

in (4.36), we obtain

$$\begin{aligned}\Omega(\zeta) = \Omega(0) + \Omega'(0)&\left\{\tilde{a}_k u^{-2/3} + 2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} u^{-1}\right\}\\ &+\frac{1}{2}\Omega''(0)\left\{\tilde{a}_k u^{-2/3} + 2^{-2/3}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1/3} u^{-1}\right\}^2 + O(u^{-2}).\end{aligned} \tag{4.45}$$

Inserting (4.37) and (4.43) in (4.45), we get

$$\begin{aligned}\Omega(\zeta) = 1 + 2^{-1/3}&\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-2/3}\tilde{a}_k u^{-2/3} + \frac{1}{2}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1} u^{-1}\\ &-\frac{\sqrt{2}}{10}\left[\frac{A_n(\lambda)}{\lambda^{2m-1}} + \frac{4A_n'(\lambda)}{\lambda^{2m-2}}\right]\left[\frac{\sqrt{2}A_n(\lambda)}{\lambda^{2m-1}}\right]^{-7/3}\tilde{a}_k^2 u^{-4/3} + O(u^{-5/3}).\end{aligned} \tag{4.46}$$

When $m = 2$, (4.46) and (4.35) together give the four-term expansion

$$\begin{aligned}x_{n,k} = u^{1/4} + 2^{-1/3}&\left(\frac{A_n(\lambda)}{\lambda^3}\right)^{-2/3}\tilde{a}_k u^{-5/12} + \frac{1}{2}\left(\frac{A_n(\lambda)}{\lambda^3}\right)^{-1} u^{-9/12}\\ &-\frac{\sqrt{2}}{10}\left[\frac{A_n(\lambda)}{\lambda^3} + \frac{4A_n'(\lambda)}{\lambda^2}\right]\left[\frac{\sqrt{2}A_n(\lambda)}{\lambda^3}\right]^{-7/3}\tilde{a}_k^2 u^{-13/12} + O(u^{-15/12}).\end{aligned} \tag{4.47}$$

From (2.33) and the equation following (2.40),

$$\frac{A_n(\lambda)}{\lambda^{2m-1}} = \sum_{k=1}^{m} k^2 v_k 2^{-2k+1}\binom{2k}{k} u^{-1+k/m} + O(u^{-1}) \tag{4.48}$$

and

$$\frac{A_n'(\lambda)}{\lambda^{2m-2}} = \frac{4}{3}\sum_{k=2}^{m} k^2(k-1) v_k 2^{-2k+1}\binom{2k}{k} u^{-1+k/m} + O(u^{-1}), \tag{4.49}$$

where use has been made of the following two combinatorial identities:

$$\sum_{l=0}^{k-1} 2^{-2l}\binom{2l}{l} = 2^{-2k+1} k\binom{2k}{k}, \tag{4.50}$$

$$\sum_{l=0}^{k-1}(k-l)2^{-2l}\binom{2l}{l} = \frac{2}{3}2^{-2k-1}(k+1)k\binom{2k+2}{k+1}, \tag{4.51}$$

which can be proved by using induction. When $m = 2$, (4.48) and (4.49) simplify to

$$\frac{A_n(\lambda)}{\lambda^3} = 3 + v_1 u^{-1/2} + O(u^{-1}) \tag{4.52}$$

and

$$\frac{A_n'(\lambda)}{\lambda^2} = 4 + O(u^{-1}). \tag{4.53}$$

Inserting (4.52) and (4.53) into (4.47), we obtain

$$\begin{aligned} x_{n,k} = u^{1/4} &+ \frac{\tilde{a}_k}{18^{1/3}} u^{-5/12} + \frac{1}{6} u^{-9/12} - \frac{1}{9}\left(\frac{2}{3}\right)^{2/3} v_1 u^{-11/12} \\ &- \frac{19\tilde{a}_k^2}{90 \cdot 2^{2/3} \cdot 3^{1/3}} u^{-13/12} + O(u^{-15/12}). \end{aligned} \tag{4.54}$$

By (2.26), $u = 2^{2m} L_m n[1 + O(n^{-2})] = \frac{4}{3} n[1 + O(n^{-2})]$ when $v(x) = x^{2m}$. Hence, if $v_1 = 0$, then (4.54) agrees with the formula recently given by Bo and Wong [2], except for the order estimate of the remainder.

When $m \geq 3$, $1/2m - 4/3 \leq -(1 + 1/2m)$. Thus (4.35) and (4.46) give only a three-term expansion

$$\begin{aligned} x_{n,k} = u^{1/2m} &+ 2^{-1/3} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-2/3} \tilde{a}_k u^{-2/3+1/2m} \\ &+ \frac{1}{2} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1} u^{-1+1/2m} + O(u^{-1-1/2m}). \end{aligned} \tag{4.55}$$

Note that $\lambda = 2a_n$, $u = (2a_n)^{2m}$, and the quantity X_n in (1.15) satisfies

$$X_n \approx \lambda + \frac{1}{2A_n(\lambda)} = u^{1/2m} + \frac{1}{2}\left(\frac{A_n(\lambda)}{\lambda^{2m-1}}\right)^{-1} u^{-1+1/2m}. \tag{4.56}$$

Since $\tilde{a}_k = -i_k/3^{1/3}$, our result (4.55) indeed agrees with (1.15) and (1.16) conjectured by Chen and Ismail [3].

Substituting (4.48) into (4.55), we obtain the second main result of this paper.

THEOREM 4.1. *Let the zeros of the polynomial $p_n(x)$ be enumerated in decreasing order: $-\infty < x_{n,n} < \cdots < x_{n,2} < x_{n,1} < \infty$, and let $u = (2a_n)^{2m}$. For each fixed k, we have*

$$x_{n,k} = u^{1/2m} + 2^{-1/3} \sigma_n(u)^{-2/3} \tilde{a}_k u^{-2/3+1/2m} + \frac{1}{2} \sigma_n(u)^{-1} u^{-1+1/2m} + O(u^{-1-1/2m}),$$

where

$$\sigma_n(u) = \sum_{k=1}^{m} k^2 v_k 2^{-2k+1} \binom{2k}{k} u^{-1+k/m}$$

and

$$u = 2^{2m} L_m n \left[1 + \sum_{k=1}^{m} \tau_k n^{-k/m} + O(n^{-1-1/m})\right]. \tag{4.57}$$

The coefficients τ_k in (4.57) can be determined by (3.23). Furthermore, if $m = 2$, then we have the five-term expansion (4.54).

5. The case $v(x) = x^{2m}$. In this section, we consider only the special case $v(x) = x^{2m}$. Although special, this is an important case. Here we observe that (2.22) and (2.23) simplify to

$$A_n(x) = 2a_n m \sum_{l=0}^{m-1} x^{2m-2l-2} c_{n,0,2l} \tag{5.1}$$

and

$$B_n(x) = 2a_n m \sum_{l=0}^{m-2} x^{2m-2l-3} c_{n,1,2l+1}. \tag{5.2}$$

Thus, instead of (2.24), we have (2.26):

$$a_n = (L_m n)^{1/2m}\left[1 + \frac{\eta}{n^2} + O\left(\frac{1}{n^4}\right)\right], \qquad n \to \infty. \tag{5.3}$$

Let $\lambda = 2a_n$ and $u = \lambda^{2m}$. Then

$$u = 2^{2m} L_m n\left[1 + \frac{2m\eta}{n^2} + O\left(\frac{1}{n^4}\right)\right]. \tag{5.4}$$

Note that the order estimate in (5.3) is much better than that in (2.24). As a result, (2.29) becomes

$$a_{n-j} = a_n\left[1 - \frac{j}{2mn} + O\left(\frac{1}{n^2}\right)\right], \tag{5.5}$$

and (2.30) becomes

$$c_{n,k,l} = \binom{l}{\frac{1}{2}(k+l)} a_n^l\left[1 + \frac{\tilde{c}_{k,l}}{2mn} + O\left(\frac{1}{n^2}\right)\right]. \tag{5.6}$$

Furthermore, all capital Os in sections 2–4 such as

$$O\left(\frac{1}{n^{1+1/m}}\right), \quad O\left(\frac{1}{\lambda^{2m+2}}\right), \quad \text{and} \quad O\left(\frac{1}{u^{1+1/m}}\right)$$

are to be replaced by

$$O\left(\frac{1}{n^2}\right), \qquad O\left(\frac{1}{\lambda^{4m}}\right), \qquad \text{and} \qquad O\left(\frac{1}{u^2}\right),$$

respectively.

Because of these changes, we now have

$$A_n(\lambda w) = m\lambda^{2m-1} \sum_{l=0}^{m-1} w^{2m-2l-2} 2^{-2l} \binom{2l}{l}\left[1 + \frac{\tilde{d}_{2l}}{\lambda^{2m}} + O\left(\frac{1}{\lambda^{4m}}\right)\right], \tag{5.7}$$

$$B_n(\lambda w) = m\lambda^{2m-1} \sum_{l=0}^{m-2} w^{2m-2l-3} 2^{-2l-1} \binom{2l+1}{l+1}\left[1 + \frac{\tilde{d}_{2l+1}}{\lambda^{2m}} + O\left(\frac{1}{\lambda^{4m}}\right)\right], \tag{5.8}$$

$$A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}} = m\lambda^{2m-1}\sum_{l=0}^{m-1} w^{2m-2l-2}2^{-2l}\binom{2l}{l}\left[1+\frac{\tilde{d}'_{2l}}{\lambda^{2m}}+O\left(\frac{1}{\lambda^{4m}}\right)\right], \tag{5.9}$$

and

$$A'_n(\lambda w) = 2m\lambda^{2m-2}\sum_{l=0}^{m-2}(m-l-1)w^{2m-2l-3}2^{-2l}\binom{2l}{l}\left[1+\frac{\tilde{d}_{2l}}{\lambda^{2m}}+O\left(\frac{1}{\lambda^{4m}}\right)\right], \tag{5.10}$$

instead of some of the equations in section 2 such as (2.33) and (2.34). Let us rewrite (5.7) in the form

$$\begin{aligned}\frac{A_n(\lambda w)}{\lambda^{2m-1}} &= m\sum_{l=0}^{m-1} w^{2m-2l-2}2^{-2l}\binom{2l}{l}\\ &\quad+\frac{2^{2m}L_m}{2}\left[\sum_{l=1}^{m-1} w^{2m-2l-2}2^{-2l}\binom{2l}{l}\tilde{c}_{0,2l}\right]u^{-1}+O(u^{-2}).\end{aligned} \tag{5.11}$$

Here use has been made of the facts that $\tilde{d}_0 = 0$ and $\tilde{d}_{2l} = 2^{2m}L_m\tilde{c}_{0,2l}/2m$. For convenience, we put

$$\bar{A}_0(w) = m\sum_{l=0}^{m-1} w^{2m-2l-2}2^{-2l}\binom{2l}{l}$$

and

$$\bar{A}_1(w) = \frac{2^{2m}L_m}{2}\left[\sum_{l=1}^{m-1} w^{2m-2l-2}2^{-2l}\binom{2l}{l}\tilde{c}_{0,2l}\right]$$

so that (5.11) becomes

$$\frac{A_n(\lambda w)}{\lambda^{2m-1}} = \bar{A}_0(w)+\bar{A}_1(w)u^{-1}+O(u^{-2}). \tag{5.12}$$

Inserting (5.12) in (2.56), and reexpanding, we get

$$\zeta = \zeta(\lambda, w) = \bar{\zeta}_0(w)+\bar{\zeta}_1(w)u^{-1}+O(u^{-2}), \tag{5.13}$$

where $\bar{\zeta}_0(w) = \zeta_\infty(w)$ is given in (2.60), that is,

$$\bar{\zeta}_0(w) = \begin{cases} \left[\frac{3}{2}m\sum_{l=0}^{m-1}2^{-2l}\binom{2l}{l}\int_1^w \sqrt{t^2-1}t^{2m-2l-2}dt\right]^{2/3}, & w \geq 1,\\ -\left[\frac{3}{2}m\sum_{l=0}^{m-1}2^{-2l}\binom{2l}{l}\int_w^1 \sqrt{1-t^2}t^{2m-2l-2}dt\right]^{2/3}, & |w| < 1,\end{cases} \tag{5.14}$$

and

$$\bar{\zeta}_1(w) = \begin{cases} \frac{2^{2m}L_m}{2\bar{\zeta}_0^{1/2}}\sum_{l=1}^{m-1}2^{-2l}\tilde{c}_{0,2l}\binom{2l}{l}\int_1^w \sqrt{t^2-1}t^{2m-2l-2}dt, & w \geq 1,\\ -\frac{2^{2m}L_m}{2(-\bar{\zeta}_0)^{1/2}}\sum_{l=1}^{m-1}2^{-2l}\tilde{c}_{0,2l}\binom{2l}{l}\int_w^1 \sqrt{1-t^2}t^{2m-2l-2}dt, & |w| < 1.\end{cases} \tag{5.15}$$

We now return to (2.68). A combination of (2.62), (2.55), and (2.53) gives

$$\Phi(\zeta)=\begin{cases}\dfrac{1}{2\zeta^{1/2}}\displaystyle\int_1^w\frac{\lambda^{2m-1}H_1(\lambda,t)}{(t^2-1)^{1/2}A_n(\lambda t)}dt, & w>1,\\ \dfrac{1}{2(-\zeta)^{1/2}}\displaystyle\int_w^1\frac{\lambda^{2m-1}H_1(\lambda,t)}{(1-t^2)^{1/2}A_n(\lambda t)}dt, & |w|<1.\end{cases}\tag{5.16}$$

Recall that $\tilde{d'}_{2l}=\tilde{d}_{2l}-l2^{2m}L_m/m$. Hence, by (5.7) and (5.9),

$$\lambda\left[A_n(\lambda w)-A_{n-1}(\lambda w)\frac{a_n}{a_{n-1}}\right]=2^{2m}L_m\sum_{l=0}^{m-1}w^{2m-2l-2}2^{-2l}l\binom{2l}{l}+O\left(\frac{1}{\lambda^{2m}}\right).\tag{5.17}$$

Similarly, we obtain

$$\lambda w[B_n(\lambda w)-B_{n+1}(\lambda w)]=-2^{2m}L_m\sum_{l=1}^{m-1}w^{2m-2l}2^{-2l}2l\binom{2l-1}{l}+O\left(\frac{1}{\lambda^{2m}}\right);\tag{5.18}$$

cf. (2.39) and (2.40). Since $H_1(\lambda,w)=\lambda^{-2m}U_1(\lambda w,n)$, combining the last two equations with (2.47), we have

$$\frac{\lambda^{2m-1}H_1(\lambda,w)}{A_n(\lambda w)}=2^{2m}L_m(1-w^2)\sum_{l=1}^{m-1}w^{2m-2l-2}2^{-2l}l\binom{2l}{l}-1+O(u^{-1}).\tag{5.19}$$

Inserting (5.19) in (5.16) gives

$$\Phi(\zeta)=\begin{cases}-\dfrac{2^{2m}L_m}{2\zeta^{1/2}}\displaystyle\sum_{l=1}^{m-1}2^{-2l}l\binom{2l}{l}\int_1^w\sqrt{t^2-1}t^{2m-2l-2}dt-\frac{\cosh^{-1}w}{2\zeta^{1/2}}+O(u^{-1}), & w>1,\\ \dfrac{2^{2m}L_m}{2(-\zeta)^{1/2}}\displaystyle\sum_{l=1}^{m-1}2^{-2l}l\binom{2l}{l}\int_w^1\sqrt{1-t^2}t^{2m-2l-2}dt-\frac{\cos^{-1}w}{2(-\zeta)^{1/2}}+O(u^{-1}), & |w|<1.\end{cases}$$

Since $\zeta(\lambda,w)=\bar{\zeta}_0(w)+\bar{\zeta}_1(w)u^{-1}+O(u^{-2})$, it follows that

$$\Phi(\zeta)=\Phi_0(w)+O(u^{-1}),\tag{5.20}$$

where

$$\Phi_0(w)=\begin{cases}-\dfrac{2^{2m}L_m}{2\bar{\zeta}_0^{1/2}}\displaystyle\sum_{l=1}^{m-1}2^{-2l}l\binom{2l}{l}\int_1^w\sqrt{t^2-1}t^{2m-2l-2}dt-\frac{\cosh^{-1}w}{2\bar{\zeta}_0^{1/2}}, & w>1,\\ \dfrac{2^{2m}L_m}{2(-\bar{\zeta}_0)^{1/2}}\displaystyle\sum_{l=1}^{m-1}2^{-2l}l\binom{2l}{l}\int_w^1\sqrt{1-t^2}t^{2m-2l-2}dt-\frac{\cos^{-1}w}{2(-\bar{\zeta}_0)^{1/2}}, & |w|<1.\end{cases}$$

If we let

$$\bar{\Phi}_0(w)=\Phi_0(w)+\bar{\zeta}_1(w),\tag{5.21}$$

then

$$X(\zeta) = u^{2/3}\zeta + \frac{\Phi(\zeta)}{u^{1/3}} = u^{2/3}\bar{\zeta}_0 + \frac{\bar{\Phi}_0(w)}{u^{1/3}} + O(u^{-4/3}). \tag{5.22}$$

By the mean-value theorem,

$$\mathrm{Ai}(X(\zeta)) = \mathrm{Ai}\left(u^{2/3}\bar{\zeta}_0 + \frac{\bar{\Phi}_0}{u^{1/3}}\right) + O(u^{-4/3}). \tag{5.23}$$

Using (5.3) instead of (2.24), the results in (3.8) and (3.38) can be improved to read

$$p_n(0) = (-1)^{n/2} A_2 n^{-\frac{1}{4m}} [1 + O(n^{-1})]$$

and

$$C(n) = A n^{-1/4m} u^{1/4m-1/3} [1 + O(u^{-1})],$$

respectively. Therefore, (3.55) becomes

$$\begin{aligned} p_n(x)e^{-v(x)/2} = L_m^{-1/4m} n^{-1/4m} u^{1/6} \left(\frac{\zeta}{w^2-1}\right)^{1/4} \{&\mathrm{Ai}(X(\zeta)) \\ &+ O(\varepsilon_1(\lambda,\zeta))\}[1 + O(u^{-1})]. \end{aligned} \tag{5.24}$$

Since $\varepsilon_1(\lambda,\zeta) = O(u^{-1})$ uniformly for $w \in [-1+\varepsilon, M]$, we get

$$p_n(x)e^{-v(x)/2} = L_m^{-1/4m} n^{-1/4m} u^{1/6} \left(\frac{\zeta}{w^2-1}\right)^{1/4} \{\mathrm{Ai}(X(\zeta)) + O(u^{-1})\}. \tag{5.25}$$

When $w \in [-1+\varepsilon, 1-\varepsilon]$, we have the sharper estimate $\varepsilon_1(\lambda,\zeta) = O(u^{-7/6})$. Since $X(\zeta) = u^{2/3}\zeta(1 + \Phi(\zeta)/u) = O(u^{2/3})$ when $\zeta \in [\zeta_\infty(-1+\varepsilon), \zeta_\infty(1-\varepsilon)]$, we also have $\mathrm{Ai}(X(\zeta)) = O(u^{-1/6})$. Substituting these into (5.24) gives

$$p_n(x)e^{-v(x)/2} = L_m^{-1/4m} n^{-1/4m} u^{1/6} \left(\frac{\zeta}{w^2-1}\right)^{1/4} \{\mathrm{Ai}(X(\zeta)) + O(u^{-7/6})\} \tag{5.26}$$

uniformly for $w \in [-1+\varepsilon, 1-\varepsilon]$.

Inserting (5.4), (5.13), and (5.23) in (5.25) and (5.26), we obtain the following theorem.

THEOREM 5.1. *When $v(x) = x^{2m}$, the asymptotic formula*

$$\begin{aligned} p_n(x)e^{-x^{2m}/2} &= 2^{m/3} L_m^{1/6-1/4m} n^{1/6-1/4m} \left(\frac{\bar{\zeta}_0}{w^2-1}\right)^{1/4} \\ &\quad \cdot \left\{\mathrm{Ai}\left(u^{2/3}\bar{\zeta}_0 + \frac{\bar{\Phi}_0}{u^{1/3}}\right) + O(u^{-1})\right\} \end{aligned} \tag{5.27}$$

holds uniformly for $w \in [-1+\varepsilon, M]$. Furthermore, we have

$$\begin{aligned} p_n(x)e^{-x^{2m}/2} &= 2^{m/3} L_m^{1/6-1/4m} n^{1/6-1/4m} \left(\frac{\bar{\zeta}_0}{w^2-1}\right)^{1/4} \\ &\quad \cdot \left\{\mathrm{Ai}\left(u^{2/3}\bar{\zeta}_0 + \frac{\bar{\Phi}_0}{u^{1/3}}\right) + O(u^{-7/6})\right\} \end{aligned} \tag{5.28}$$

holding uniformly for $w \in [-1+\varepsilon, 1-\varepsilon]$. Here $\bar{\zeta}_0$ and $\bar{\Phi}_0$ are defined, respectively, in (5.14) *and* (5.21)*, and u is given in* (5.4).

Note that $\tilde{c}_{0,2} = 1$, and hence that when $m = 2$,

$$\bar{\Phi}_0(w) = \begin{cases} -\dfrac{1}{2\bar{\zeta}_0^{1/2}} \cosh^{-1} w, & w > 1, \\ -\dfrac{1}{2(-\bar{\zeta}_0)^{1/2}} \cos^{-1} w, & |w| < 1. \end{cases}$$

Thus we can easily check that our result coincides with that given by Bo and Wong [2].

To conclude the paper, we consider the zeros of $p_n(x)$ when $v(x) = x^{2m}$. Putting $w = 1$ in (5.19), we have

$$\frac{\lambda^{2m-1} H_1(\lambda, 1)}{A_n(\lambda)} = -1 + O(u^{-1}).$$

Comparing it with (4.32), we obtain

$$\Phi(0) = -2^{-2/3} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}} \right)^{-1/3} + O(\lambda^{-2m});$$

cf. (4.33). As in section 4, we insert it in (4.24) to get

$$\zeta_{n,k} = \tilde{a}_k u^{-2/3} + 2^{-2/3} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}} \right)^{-1/3} u^{-1} + O(u^{-5/3}), \tag{5.29}$$

where $\zeta_{n,k} = \zeta(\lambda, w_{n,k})$, $x_{n,k} = \lambda w_{n,k}$, and $x_{n,k}$ is the kth zero of $p_n(x)$. By inversion,

$$x_{n,k} = u^{1/2m} \Omega \left\{ \tilde{a}_k u^{-2/3} + 2^{-2/3} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}} \right)^{-1/3} u^{-1} \right\} + O(u^{-5/3+1/2m}); \tag{5.30}$$

cf. (4.35). Note that (4.46) always holds. Hence

$$\begin{aligned} x_{n,k} = u^{1/2m} &+ 2^{-1/3} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}} \right)^{-2/3} \tilde{a}_k u^{-2/3+1/2m} + \frac{1}{2} \left(\frac{A_n(\lambda)}{\lambda^{2m-1}} \right)^{-1} u^{-1+1/2m} \\ &- \frac{\sqrt{2}}{10} \left[\frac{A_n(\lambda)}{\lambda^{2m-1}} + \frac{4A_n'(\lambda)}{\lambda^{2m-2}} \right] \left[\frac{\sqrt{2} A_n(\lambda)}{\lambda^{2m-1}} \right]^{-7/3} \tilde{a}_k^2 u^{-4/3+1/2m} + O(u^{-5/3+1/2m}). \end{aligned} \tag{5.31}$$

Since (5.7) and (5.10) can be written as

$$\frac{A_n(\lambda w)}{\lambda^{2m-1}} = m \sum_{l=0}^{m-1} w^{2m-2l-2} 2^{-2l} \binom{2l}{l} + O\left(\frac{1}{\lambda^{2m}} \right)$$

and

$$\frac{A_n'(\lambda w)}{\lambda^{2m-2}} = 2m \sum_{l=0}^{m-2} (m-l-1) w^{2m-2l-3} 2^{-2l} \binom{2l}{l} + O\left(\frac{1}{\lambda^{2m}} \right),$$

putting $w = 1$ and using (4.50) and (4.51), we have

$$\frac{A_n(\lambda)}{\lambda^{2m-1}} = 2^{-2m+1} m L_m^{-1} + O(u^{-1}) \tag{5.32}$$

and

$$\frac{A_n'(\lambda)}{\lambda^{2m-2}} = \frac{4}{3}2^{-2m+1}m(m-1)L_m^{-1} + O(u^{-1}), \tag{5.33}$$

where we have also made use of (2.25). Substituting (5.32) and (5.33) into (5.31), we obtain

$$\begin{aligned} x_{n,k} &= u^{1/2m} + \frac{2^{(4m-1)/3}}{(2m)^{2/3}}L_m^{2/3}\tilde{a}_k u^{-2/3+1/2m} + \frac{2^{2m}L_m}{4m}u^{-1+1/2m} \\ &\quad - \frac{16m-13}{120}\frac{2^{8m/3}}{m^{4/3}}L_m^{4/3}\tilde{a}_n^2 u^{-4/3+1/2m} + O(u^{-5/3+1/2m}). \end{aligned} \tag{5.34}$$

By (5.4), (5.34) can be rewritten in the form

$$\begin{aligned} x_{n,k} &= 2(L_m n)^{1/2m} + \frac{L_m^{1/2m}}{m^{2/3}}\tilde{a}_k n^{-2/3+1/2m} + \frac{L_m^{1/2m}}{2m}n^{-1+1/2m} \\ &\quad - \frac{(16m-13)L_m^{1/2m}}{60m^{4/3}}\tilde{a}_n^2 n^{-4/3+1/2m} + O(n^{-5/3+1/2m}). \end{aligned} \tag{5.35}$$

REFERENCES

[1] W. C. Bauldry, A. Máté, and P. Nevai, *Asymptotics for solutions of systems of smooth recurrence equations*, Pacific J. Math., 133 (1988), pp. 209–227.

[2] R. Bo and R. Wong, *A uniform asymptotic formula for orthogonal polynomials associated with* $\exp(-x^4)$, J. Approx. Theory, 98 (1999), pp. 146–166.

[3] Y. Chen and E. H. Ismail, *Ladder operators and differential equations for orthogonal polynomials*, J. Phys. A., 30 (1997), pp. 7818–7829.

[4] A. Máté, P. Nevai, and V. Totik, *Asymptotics for the greatest zeros of orthogonal polynomials*, SIAM J. Math. Anal., 17 (1986), pp. 745–751.

[5] A. Máté, P. Nevai, and T. Zaslavsky, *Asymptotic expansions of ratios of coefficients of orthogonal polynomials with exponential weights*, Trans. Amer. Math. Soc., 287 (1985), pp. 495–505.

[6] P. Nevai, *Asymptotics for orthogonal polynomials associated with* $\exp(-x^4)$, SIAM J. Math. Anal., 15 (1984), pp. 1177–1187.

[7] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[8] R. C. Sheen, *Asymptotics for orthogonal polynomials associated with* $\exp(-x^6)$, J. Approx. Theory, 50 (1987), pp. 232–293.

[9] G. Szegö, *Orthogonal Polynomials*, 4th ed., Colloquium Publications, Vol. 23, AMS, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 31, No. 5, pp. 1030–1061

NONLINEAR PYRAMID TRANSFORMS BASED ON MEDIAN-INTERPOLATION*

DAVID L. DONOHO[†] AND THOMAS P.-Y. YU[‡]

Abstract. We introduce a nonlinear refinement subdivision scheme based on *median-interpolation*. The scheme constructs a polynomial interpolating adjacent block medians of an underlying object. The interpolating polynomial is then used to impute block medians at the next finer triadic scale. Perhaps surprisingly, expressions for the refinement operator can be obtained in closed-form for the scheme interpolating by polynomials of degree $D = 2$. Despite the nonlinearity of this scheme, convergence and regularity can be established using techniques reminiscent of those developed in analysis of linear refinement schemes.

The refinement scheme can be deployed in multiresolution fashion to construct a nonlinear pyramid and an associated forward and inverse transform. In this paper we discuss the basic properties of these transforms and their possible use in removing badly non-Gaussian noise. Analytic and computational results are presented to show that in the presence of highly non-Gaussian noise, the coefficients of the nonlinear transform have much better properties than traditional wavelet coefficients.

Key words. pyramid transform, subdivision scheme, wavelet, median, robust statistics, nonlinear analysis, interpolation

AMS subject classifications. 26A15, 26A16, 26A18, 26A27, 39B05, 41A05, 41A46, 42C40, 94A12, 62G05, 62G35

PII. S0036141097330294

1. Introduction.

Recent theoretical studies [14, 13] have found that the orthogonal wavelet transform offers a promising approach to noise removal. They assume that one has noisy samples of an underlying function f

$$y_i = f(t_i) + \sigma z_i,\ i = 1, \dots, n, \tag{1.1}$$

where $(z_i)_{i=1}^n$ is a standard Gaussian white noise and σ is the noise level. In this setting, they show that one removes noise successfully by applying a wavelet transform, thresholding the wavelet coefficients, and inverting the transform. Here "success" means near-asymptotic minimaxity over a broad range of classes of smooth f. Other efforts [20, 17] have shown that the Gaussian noise assumption can be relaxed slightly; in the presence of non-Gaussian noise that is not too heavy-tailed (e.g., the density has sufficiently rapid decay at $\pm\infty$), one can use level-dependent thresholds which are somewhat higher than in the Gaussian case and continue to obtain near-minimax results.

1.1. Strongly non-Gaussian noise.

In certain settings, data exhibit strongly non-Gaussian noise distributions; examples include analogue telephony [30], radar signal processing [1], and laser radar imaging [19]. By strongly non-Gaussian we mean subject to very substantial deviations much more frequently than under Gaussian assumptions.

*Received by the editors November 19, 1997; accepted for publication (in revised form) April 2, 1999; published electronically April 20, 2000. This research was partially supported by NSF DMS-95-05151, by AFOSR MURI95-F49620-96-1-0028, and by other sponsors.

http://www.siam.org/journals/sima/31-5/33029.html

[†] Department of Statistics, Stanford University, Stanford, CA 94305 (donoho@stat.stanford.edu).

[‡] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (yut@rpi.edu).

Thresholding of linear wavelet transforms *does not* work well with strongly non-Gaussian noise. Consider model (1.1) in a specific case: let (z_i) be independently and identically Cauchy distributed. The Cauchy distribution has no moments $\int x^\ell f(x)dx$ for $\ell = 1, 2, \ldots$, in particular neither mean nor variance.

Under this model, typical noise realizations $(z_i)_{i=1}^n$ contain a few astonishingly large observations: the largest observation is of size $O(n)$. (In comparison, for Gaussian noise, the largest observation is of size $O(\sqrt{\log(n)})$.) Moreover, a linear wavelet transform of independently and identically distributed (i.i.d.) Cauchy noise does not result in independent, nor identically distributed wavelet coefficients. In fact, coefficients at coarser scales are more likely to be affected by the perturbing influence of the few large noise values, and so one sees a systematically larger stochastic dispersion of coefficients at coarse scales. Invariance of distribution across scale and $O(\sqrt{\log(n)})$ behavior of maxima are fundamental to the results on wavelet denoising in [13, 14]. The Cauchy situation therefore lacks key quantitative properties which were used in denoising in the Gaussian case.

It is not just that this situation lacks properties which would make the proofs "go through." If we try to apply ideas which were successful under Gaussian theory we meet with abject failure, as simple computational examples given later will illustrate.

1.2. Median-interpolating pyramid transform. Motivated by this situation, this paper develops a kind of *nonlinear "wavelet transform."* The need to abandon linearity is clear a priori. It is well known that linearity of approach is essentially tantamount to a Gaussian assumption and that non-Gaussian assumptions typically lead to highly nonlinear approaches. For example, maximum likelihood estimators in classical statistical models are often linear in the Gaussian case, but highly nonlinear under Cauchy and similar assumptions.

Central to our approach is the notion of *median-interpolating* (MI) refinement scheme. Given data about the medians of an object on triadic blocks at a coarse scale, we predict the medians of triadic blocks at the next finer scale. We do this by finding a median-interpolating polynomial—a polynomial with the same coarse-scale block medians—and then calculating the block medians of this polynomial for blocks at the next finer scale. The procedure is nonlinear: the interpolating polynomial is a nonlinear functional of the coarse-scale medians; and the imputed finer-scale medians are nonlinear functionals of the interpolating polynomial. Perhaps surprisingly, in the case of interpolation by quadratic polynomials, the interpolating polynomial and its finer-scale medians can both be found in closed-form.

Using MI refinement, we can build forward and inverse transforms which can be computed rapidly and which exhibit favorable robustness and regularity properties.

The forward transform deploys the median in a multiresolution pyramid; it computes "block medians" over all triadic cells. MI refinement is used to predict medians of finer-scale blocks from coarser-scale blocks; the resulting prediction errors are recorded as transform coefficients.

The inverse transform undoes this process; using coarser-scale coefficients it builds MI predictions of finer-scale coefficients; adding in the prediction errors recorded in the transform array leads to exact reconstruction.

This way of building transforms from refinement schemes is similar to the way interpolating wavelet transforms are built from Deslauriers–Dubuc interpolating schemes in [9] and in the way biorthogonal wavelet transforms are built from average-interpolating refinement in [10]. The basic idea is to use data at coarser scales to predict data at finer scales and to record the prediction errors as coefficients associated with the finer scales. Despite structural similarities our MI-based transforms exhibit important differences:

- Both the forward and inverse transforms can be nonlinear;
- The transforms are based on a triadic pyramid and a 3-to-1 decimation scheme;
- The transforms are expansive (they map n data into $\sim 3/2n$ coefficients).

Terminologically, because the forward transform is expansive, it should be called a *pyramid transform* rather than a wavelet transform. We call the transform itself the *median-interpolating pyramid transform* (MIPT).

The bulk of our paper is devoted to analysis establishing two key properties of these transforms.

- *Regularity.* Take block medians at a single level and refine to successively finer and finer levels using the quadratic polynomial MI scheme. Detailed analysis shows that the successive refinements converge uniformly to a continuous limit with Hölder-α regularity for some $\alpha > 0$. We prove that $\alpha > .0997$ and we give computational and analytical evidence pointing to $\alpha > 1 - \epsilon$ for all $\epsilon > 0$.

  This result shows that MIPT has important similarities to linear wavelet and pyramid transforms. For example, it provides a notion of nonlinear multiresolution analysis: just as in the linear MRA case, one can decompose an object into "resolution levels" and examine the contributions of different levels separately; each level contributes a regular curve oscillating with wavelength comparable to the given resolution, with large oscillations in the spatial vicinity of significant features.
- *Robustness.* It is well known that the median is robust against heavy-tailed noise distributions [21, 22]. In the present setting this phenomenon registers as follows. We are able to derive thresholds for noise removal in the MIPT which work well for *all* distributions in rather large classes, irrespective of the heaviness of the tails. In particular, we show that at all but the finest scales, the same thresholds work for both Gaussian and Cauchy data. Hence a noise-removal scheme based on thresholding of MIPT coefficients depends only very weakly on assumptions about noise distribution.

There is considerable applied interest in developing median-based multiscale transforms, as one can see from [2, 23, 28, 29, 26]. The analysis we give here suggests that our framework will turn out to have strong theoretical justification and may provide applied workers with helpful new tools.

1.3. Contents.

Section 2 introduces the notion of median-interpolating refinement, shows how one of the simplest instances may be computed efficiently, gives computational examples, and proves some basic properties. Section 3 establishes convergence and smoothness results for the quadratic median-interpolating refinement scheme. Section 4 develops a nonlinear pyramid transform and describes properties of transform coefficients. Proofs of these properties are recorded in section 6. Section 5 applies the pyramid transform to the problem of removing highly non-Gaussian noise.

2. Median-interpolating refinement schemes.

In this section we describe a notion of two-scale refinement which is nonlinear in general, and which yields an interesting analogue of the refinement schemes occurring in the theory of biorthogonal wavelets.

2.1. Median-interpolation. Given a function f on an interval I, let $\mathrm{med}(f|I)$ denote a median of f for the interval I, defined by

$$\mathrm{med}(f|I) = \inf\{\mu : m(t \in I : f(t) \geq \mu) \geq m(t \in I : f(t) \leq \mu)\}, \tag{2.1}$$

where $m()$ denotes Lebesgue measure on $\mathbb{R}$.

Now suppose we are given a triadic array $\{m_{j,k}\}_{k=0}^{3^j-1}$ of numbers representing the medians of f on the triadic intervals $I_{j,k} = [k3^{-j}, (k+1)3^{-j})$:

$$m_{j,k} = \mathrm{med}(f|I_{j,k}), \quad 0 \leq k < 3^j, \quad j \geq 0.$$

The goal of median-interpolating refinement is to use the data at scale j to infer behavior at the finer scale $j+1$, obtaining imputed medians of f on intervals $I_{j+1,k}$. Obviously we are missing the information to impute perfectly; nevertheless we can try to do a reasonable job.

We employ *polynomial-imputation.* Starting from a fixed even integer D, it involves two steps.

[M1] **(Interpolation).** For each interval $I_{j,k}$, find a polynomial $\pi_{j,k}$ of degree $D = 2A$ satisfying the *median-interpolation condition*:

$$\mathrm{med}(\pi_{j,k}|I_{j,k+l}) = m_{j,k+l} \quad \text{for} \quad -A \leq l \leq A. \tag{2.2}$$

[M2] **(Imputation).** Obtain (pseudo-) medians at the finer scale by setting

$$\tilde{m}_{j+1,3k+l} = \mathrm{med}(\pi_{j,k}|I_{j,3k+l}) \quad \text{for} \quad l = 0, 1, 2. \tag{2.3}$$

An example is given in Figure 2.1 for degree $D = 2$. Some questions come up naturally:

[Q1] Is there a unique polynomial $\pi_{j,k}$ satisfying the nonlinear equations (2.2)?
[Q2] If so, is there an effective algorithm to find it?
[Q3] If so, what are the properties of such a procedure?

2.1.1. Average-interpolation. A scheme similar to the above, with "med" replaced by "ave," is relatively easy to study and provides useful background. Given a function f on an interval I, write $\mathrm{ave}(f|I) = |I|^{-1} \int_I f(t)dt$ for the average value of f over the interval I. Now suppose we are given a triadic array $\{a_{j,k}\}_{k=0}^{3^j-1}$ of numbers representing the averages of f on the triadic intervals $I_{j,k}$. Average-interpolating refinement uses the data at scale j to impute behavior at the finer scale $j+1$, obtaining the (pseudo-) averages of f on intervals $I_{j+1,k}$. Fix an even integer D, it runs as follows:

[A1] **(Interpolation).** For each interval $I_{j,k}$, find a polynomial $\pi_{j,k}$ of degree $D = 2A$ satisfying the *average-interpolation condition*:

$$\mathrm{ave}(\pi_{j,k}|I_{j,k+l}) = a_{j,k+l} \quad \text{for} \quad -A \leq l \leq A. \tag{2.4}$$

[A2] **(Imputation).** Obtain (pseudo-) cell averages at the finer scale by setting

$$\overline{a}_{j+1,3k+l} = \mathrm{ave}(\pi_{j,k}|I_{j,3k+l}) \quad \text{for } l = 0, 1, 2. \tag{2.5}$$

This type of procedure has been implemented and studied in (the dyadic case) [10, 11]. The analogues of questions [Q1]–[Q2] have straightforward "Yes" answers. For any degree D one can find coefficients $c_{h,l}^{(D)}$ for which

$$\overline{a}_{j+1,3k+l} = \sum_{h=-A}^{A} c_{h,l}^{(D)} a_{j,k+h}, \quad l = 0, 1, 2, \tag{2.6}$$

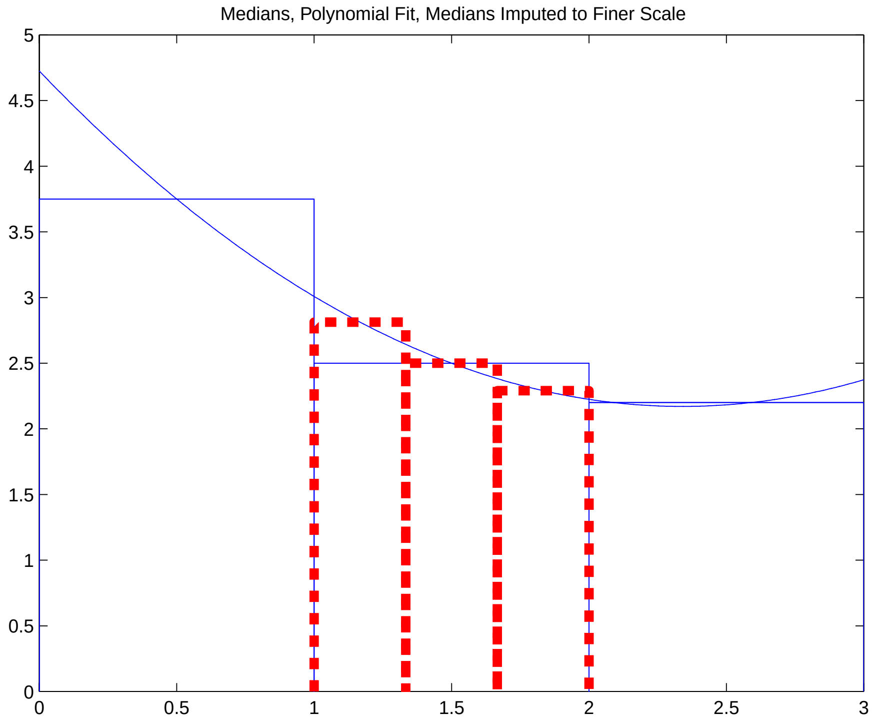


FIG. 2.1. *Median-interpolation, $D = 2$. The rectangular blocks surrounded by solid lines correspond to $m_{0,0}, m_{0,1}, m_{0,2}$, the blocks surrounded by thickened dashed lines correspond to imputed medians $\tilde{m}_{1,2}, \tilde{m}_{1,3}, \tilde{m}_{1,4}$, the parabola corresponds to the median-interpolant $\pi_{0,1}$.*

exhibiting the fine-scale imputed averages $\overline{a}_{j+1,k}$'s as linear functionals of the coarse-scale averages $a_{j,k}$. Moreover, using analytic tools developed in wavelet theory [4] and in refinement subdivision schemes [8, 16] one can establish various nice properties of refinement by average-interpolation—see below.

2.1.2. $D = 0$. We return to median-interpolation. The case $D = 0$ is the simplest by far; in that case one is fitting a constant function $\pi_{j,k}(t) = \text{Const}$. Hence $A = 0$, and (2.2) becomes $\pi_{j,k}(t) = m_{j,k}$. The imputation step (2.3) then yields $m_{j+1,3k+l} = m_{j,k}$ for $l = 0, 1, 2$. Hence refinement proceeds by imputing a constant behavior at finer scales.

2.1.3. $D = 2$. The next simplest case is $D = 2$ and will be the focus of attention in this article. To apply (2.2) with $A = 1$, we must find a quadratic polynomial solving

$$\text{med}(\pi_{j,k}|I_{j,k+l}) = m_{j,k+l} \quad \text{for } l = -1, 0, 1. \tag{2.7}$$

In general this is a system of nonlinear equations. One can ask [Q1]–[Q3] above for this system. The answers come by studying the operator $\Pi_{(2)} : \mathbb{R}^3 \to \mathbb{R}^3$ defined as the solution to the problem: given $[m_1, m_2, m_3]$, find $[a, b, c]$ such that the quadratic polynomial $\pi(x) = a + bx + cx^2$ satisfies

$$\text{med}(\pi|[0, 1]) = m_1, \tag{2.8}$$

$$\operatorname{med}(\pi|[1,2]) = m_2, \tag{2.9}$$

$$\operatorname{med}(\pi|[2,3]) = m_3. \tag{2.10}$$

In this section, we work out explicit algebraic formulae for $\Pi_{(2)}$. It will follow from these that (2.7) *has an unique solution, for every* m_1, m_2, m_3, and that this solution is a Lipschitz function of the m_i.

$\Pi_{(2)}$ possesses two purely formal invariance properties which are useful below.

- *Reversal equivariance.* If $\Pi_{(2)}(m_1, m_2, m_3) = a + bx + cx^2$, then $\Pi_{(2)}(m_3, m_2, m_1) = a + b(3-x) + c(3-x)^2$.
- *Affine equivariance.* If $\Pi_{(2)}(m_1, m_2, m_3) = \pi$, then $\Pi_{(2)}(a + bm_1, a + bm_2, a + bm_3) = a + b\pi$.

Reversal equivariance is, of course, tied to the fact that median-interpolation is a spatially symmetric operation. From affine equivariance, it follows that when $m_2 - m_1 \neq 0$ we have

$$\begin{aligned}\Pi_{(2)}(m_1, m_2, m_3) &= m_1 + \Pi_{(2)}(0, m_2 - m_1, m_3 - m_1) \\ &= m_1 + (m_2 - m_1)\Pi_{(2)}\left(0, 1, 1 + \frac{m_3 - m_2}{m_2 - m_1}\right).\end{aligned} \tag{2.11}$$

Thus $\Pi_{(2)}$ is characterized by its action on very special triples; it is enough to study the univariate function $\Pi_{(2)}(0, 1, 1 + d)$, $d \in \mathbb{R}$. (The exceptional case when $m_2 - m_1 = 0$ can be handled easily; see the discussion after the proof of Proposition 2.2.)

To translate (2.8)–(2.10) into manageable algebraic equations, we begin with the following proposition.

PROPOSITION 2.1 (median-imputation, $D = 2$). *Suppose the quadratic polynomial* $\pi(x)$ *has its extremum at* x^*. *Let* $s = q - p$.

[L] *If* $x^* \notin [p + s/4, p + 3s/4]$, *then*

$$\operatorname{med}(\pi(x)|[p, q]) = \pi((p + q)/2). \tag{2.12}$$

[N] *If* $x^* \in [p + s/4, p + 3s/4]$, *then*

$$\operatorname{med}(\pi(x)|[p, q]) = \pi(x^* \pm s/4). \tag{2.13}$$

Proof. We assume x^* is a minimizer (the case of a maximizer being similar). The key fact is that $\pi(x)$, being a quadratic polynomial, is symmetric about x^* and monotone increasing in $|x - x^*|$.

If $x^* \in [p + s/4, p + 3s/4]$, then $[x^* - s/4, x^* + s/4] \subseteq [p, q]$, $\{x \in [p, q] \mid \pi(x) \leq \pi(x^* \pm s/4)\} = [x^* - s/4, x^* + s/4]$. Thus $m\{x \in [p, q] \mid \pi(x) \leq \pi(x^* \pm s/4)\} = s/2 = m\{x \in [p, q] \mid \pi(x) \geq \pi(x^* \pm s/4)\}$, which implies $\operatorname{med}(\pi(x)|[p, q]) = \pi(x^* \pm s/4)$.

If $x^* < p + s/4$, then $\{x \in [p, q] \mid \pi(x) \leq \pi((p + q)/2)\} = [p, p + s/2]$ and $\{x \in [p, q] \mid \pi(x) \geq \pi((p + q)/2)\} = [p + s/2, q]$ have equal measure. Thus $\operatorname{med}(\pi(x)|[p, q]) = \pi((p + q)/2)$. The same conclusion holds when $x^* > p + 3s/4$. Thus we have (2.12). ▯

The two cases identified above will be called the "Linear" and "Nonlinear" cases. Equations (2.8)–(2.10) always give rise to a system of three algebraic equations in three variables a, b, c. Linearity refers to dependence on these variables. When (2.13) is invoked, $x^* = -b/2c$, and so the evaluation of π—at a location depending on x^*—is a nonlinear functional.

A similar division into cases occurs when we consider median-interpolation.

PROPOSITION 2.2 (median-interpolation, $D = 2$). $\Pi_{(2)}(0, 1, 1 + d) = a + bx + cx^2$ *can be computed by the following formulae:*

[N1] *If* $\frac{7}{3} \le d \le 5$, *then* $x^* \in [\frac{1}{4}, \frac{3}{4}]$, *and*

$$a = 11 + \frac{7}{2}d - \frac{5}{2}r, \;\; b = -\frac{32}{3} - \frac{13}{3}d + \frac{8}{3}r, \;\; c = \frac{8}{3} + \frac{4}{3}d - \frac{2}{3}r, \tag{2.14}$$

where $r = \sqrt{16 + 16d + d^2}$.

[N2] *If* $\frac{1}{5} \le d \le \frac{3}{7}$, *then* $x^* \in [\frac{9}{4}, \frac{11}{4}]$, *and*

$$a = -\frac{3}{2} - 2d + \frac{1}{2}d - \frac{5}{2}r, \; b = -\frac{11}{3} + \frac{16}{3}d - \frac{4}{3}r, \; c = -\frac{4}{3} - \frac{8}{3}d + \frac{2}{3}r, \tag{2.15}$$

where $r = \sqrt{1 + 16d + 16d^2}$.

[N3] *If* $-3 \le d \le -\frac{1}{3}$, *then* $x^* \in [\frac{5}{4}, \frac{7}{4}]$, *and*

$$a = -\frac{7}{12} + \frac{1}{12}d + \frac{r}{12}, \;\; b = \frac{13}{10} - \frac{3}{10}d - \frac{r}{5}, \;\; c = -\frac{4}{15} + \frac{4}{15}d + \frac{r}{15}, \tag{2.16}$$

where $r = -\sqrt{1 - 62d + d^2}$.

[L] *In all other cases,*

$$a = -\frac{7}{8} + \frac{3}{8}d, \;\; b = 2 - d, \;\; c = -\frac{1}{2} + \frac{d}{2}. \tag{2.17}$$

Proof. Fix a polynomial π. To calculate its block medians on blocks $[0, 1]$, $[1, 2]$, $[2, 3]$, we can apply Proposition 2.1 successively to the choices $[p, q] = [0, 1]$, $[1, 2]$, $[2, 3]$. We see that either the extremum of π lies in the middle half of one of the three intervals $[0, 1]$, $[1, 2]$, $[2, 3]$ or it does not. If it does not lie in the middle half of any interval, the relation of the block medians to the coefficients is linear. If it does lie in the middle half of some interval, the relation of the block medians to the coefficients will be linear in two of the blocks—those where the extremum does not lie—and nonlinear in the block where the extremum does lie. Hence there are four basic cases to consider: (i) $x^* \in [1/4, 3/4]$, (ii) $x^* \in [9/4, 11/4]$, (iii) $x^* \in [5/4, 7/4]$, and (iv) $x^* \notin [1/4, 3/4] \cup [9/4, 11/4] \cup [5/4, 7/4]$. The first three involve some form of nonlinearity; the remaining case is linear.

Now to *solve for* a polynomial π with prescribed block medians, we can see at this point that *if we knew in advance the value of* $x^*(\pi)$, we could identify one of cases (i)–(iv) as being operative. It is easy to set up for any one of these cases a system of algebraic equations defining the desired quadratic polynomial. By writing down the system explicitly and solving it, either by hand or with the assistance of an algebraic software tool, we can obtain explicit formulae for the coefficients π. This has been done for cases (i)–(iv) and results are recorded above in (2.14)–(2.17). We omit the detailed calculation.

At this point, we have identified four different cases relating polynomials to their block medians. Within a given case, the relationship between a polynomial and its block medians is one-one. However, it remains for the moment at least conceivable that for a given collection of block medians, there would be two different cases which gave the same block medians, and hence nonunique interpolation.

We are rescued by a small miracle: *with six exceptions, a given set of block medians is consistent with exactly one of the four cases.*

To understand this, note that each of the four cases, involving a hypothesis on x^*, is consistent with block medians $[0, 1, 1 + d]$ only for a special set of values of d. We now proceed to identify the set of values of d which may arise in each given case,

case by case (but out of numerical order). Starting with case (iv), we can show that *if* $x^* \notin [1/4, 3/4] \cup [9/4, 11/4] \cup [5/4, 7/4]$*, then the associated block medians* $[0, 1, 1+d]$ *must obey* $d \notin [7/3, 5] \cup [1/5, 3/7] \cup [-3, -1/3]$. As we are in case (iv), formula (2.17) gives $\Pi_{(2)}(0, 1, 1+d) = (-7/8 + 3/8d) + (2-d)x + ((d-1)/2)x^2$, and hence

$$x^* = (d-2)/(d-1). \tag{2.18}$$

By a routine calculation, case (iv) and (2.18) combine to conclude that $d \notin [-3, -1/3] \cup [1/5, 3/7] \cup [7/3, 5]$. For future reference, set $\mathcal{L} = ([-3, -1/3] \cup [1/5, 3/7] \cup [7/3, 5])^c$.

Now turn to case (i); we can show that *if* $x^* \in [1/4, 3/4]$*, then the associated block medians* $[0, 1, 1+d]$ *must obey* $d \in [7/3, 5]$. As we are in case (i), formula (2.14) applies, and

$$x^* = \frac{32 + 13d - 8\sqrt{16 + 16d + d^2}}{4(4 + 2d - \sqrt{16 + 16d + d^2})}. \tag{2.19}$$

This and $x^* \in [1/4, 3/4]$ combine to conclude that $d \in [7/3, 5]$. For future reference, set $\mathcal{N}_1 = [7/3, 5]$.

Similar calculations show that in case (ii) we have *if* $x^* \in [9/4, 11/4]$*, then the associated block medians* $[0, 1, 1+d]$ *must obey* $d \in \mathcal{N}_2 \equiv [1/5, 3/7]$. Also in case (iii) we have *if* $x^* \in [5/4, 7/4]$*, then the associated block medians* $[0, 1, 1+d]$ *must obey* $d \in \mathcal{N}_3 \equiv [-3, -1/3]$.

We now have a collection of 4 sets: $\mathcal{L}$ and $\mathcal{N}_i$, $i = 1, 2, 3$. The sets have disjoint interiors and together cover the whole range of possible values for d. For d in the interior of one of these sets, exactly one of the four cases is able to generate the block medians $[0, 1, 1+d]$. The exceptional values of d, not in the interior of one of the sets, lie in the intersection of one of the nonlinear branches $\mathcal{N}_i$ and the linear branch $\mathcal{L}$. They are $-3, -1/3, 1/5, 3/7, 7/3, 5$. Analysis "by hand" shows that at each exceptional value of d, the formula for cases (iv) and the formula for the appropriate case (i)–(iii) give identical polynomials. Formally, one sees that if $a_L(d)$, $b_L(d)$, $c_L(d)$ denote formulas from one of the expressions (2.14)–(2.17) associated with an interval immediately to the left of an exceptional value (d_E, say), and $a_R(d)$, $b_R(d)$, $c_R(d)$ denote corresponding formulas associated with the interval immediately to the right of that same exceptional value, then

$$\lim_{d \uparrow d_E} a_L(d) = \lim_{d \downarrow d_E} a_R(d)$$

and similarly for b_L, b_R, c_L, c_R.

Thus the formulas listed above *cohere globally.* Each individual formula gives $a(d)$, $b(d)$, and $c(d)$ valid on the hypothesis that x^* lies in a certain range; but because of the continuous joining at the exceptional values, the different formulas combine to produce globally monotone functions of d. See Figure 2.2. □

The degenerate case $m_2 - m_1 = 0$ can be handled as follows: (i) if $m_1 = m_2 = m_3$, then $\Pi_{(2)}(m_1, m_2, m_3) = m_1$; (ii) otherwise $m_3 - m_2 \neq 0$, then use reversal equivariance followed by the formulae in Proposition 2.2. Notice that [N1]–[N3] are nonlinear rules, whereas [L] is linear. Figure 2.2 illustrates the nonlinearity of $\Pi_{(2)}$. Panels (a)–(c) show $a(d), b(d)$, and $c(d)$, where $a(d) + b(d)x + c(d)x^2 = \Pi_{(2)}(0, 1, d+1)$. Panel (d) shows $\partial a/\partial d$. Proposition 2.2 implies that $\Pi_{(2)}$ is basically *linear* outside

$$\mathcal{N} = \left\{ [m_1, m_2, m_3] \;\middle|\; \frac{m_3 - m_2}{m_2 - m_1} \in \mathcal{N}_0 \right\}, \tag{2.20}$$

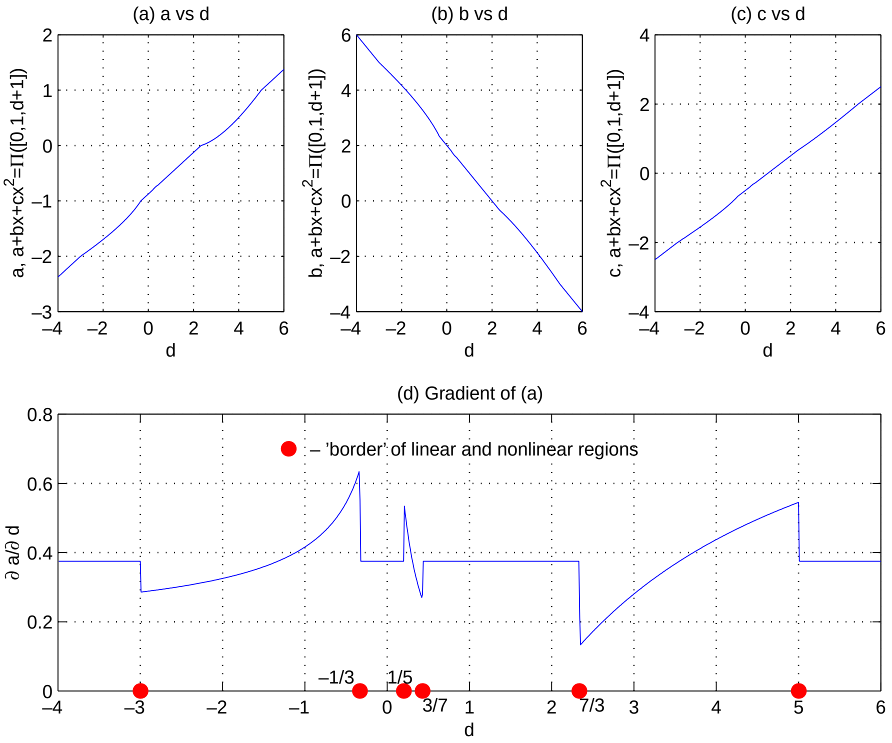


Fig. 2.2. *Nonlinearity structure of* $\Pi_{(2)}$.

where

$$\mathcal{N}_0 = [-3, -1/3] \cup [1/5, 3/7] \cup [7/3, 5]. \tag{2.21}$$

Precisely, if $\mu, \beta, a\mu + b\beta \in \mathcal{N}^c$, then

$$\Pi_{(2)}(a\mu + b\beta) = a\,\Pi_{(2)}(\mu) + b\,\Pi_{(2)}(\beta).$$

Figure 2.2(d) illustrates this point.

We now combine Propositions 2.1 and 2.2 to obtain closed-form expressions for the two-scale median-interpolating refinement operator in the quadratic case. First of all, Proposition 2.1 implies the well-posedness of median-interpolation in the case $D = 2$. Hence there exists a median-interpolating refinement operator $Q : \mathbb{R}^3 \to \mathbb{R}^3$ such that if $\pi = \pi_{j,k}$ is the fitted polynomial satisfying

$$\text{med}(\pi | I_{j,k+l}) = m_l \quad \text{for} \quad -1 \le l \le 1. \tag{2.22}$$

Then

$$Q([m_{-1}, m_0, m_1]) = [\text{med}(\pi | I_{j,3k}), \text{med}(\pi | I_{j,3k+1}), \text{med}(\pi | I_{j,3k+2})]. \tag{2.23}$$

Note that the refinement calculation is *independent* of the scale and spatial indices j and k, so Q is indeed a map from $\mathbb{R}^3$ to $\mathbb{R}^3$.

The operator Q shares two equivariance properties with $\Pi_{(2)}$:

- *Reversal equivariance.*

  $Q(m_1, m_2, m_3) = -\text{reverse}(Q(-m_3, -m_2, -m_1))$, where $\text{reverse}(p, q, r) = (r, q, p)$.

- *Affine equivariance.*

$$Q(a + bm_1, a + bm_2, a + bm_3) = a + b\,Q(m_1, m_2, m_3).$$

Q is characterized by its action on triplets $(0, 1, 1 + d)$, since if $m_2 - m_1 \neq 0$,

$$Q(m_1, m_2, m_3) = m_1 + (m_2 - m_1)Q\left(0, 1, 1 + \frac{m_3 - m_2}{m_2 - m_1}\right), \tag{2.24}$$

while if $m_2 - m_1 = 0$ and $m_3 - m_2 \neq 0$, then $Q(m_1, m_2, m_3) = \text{reverse}(Q(m_3, m_2, m_1))$ and $Q(m_3, m_2, m_1)$ can then be calculated from (2.24). Of course, when $m_2 - m_1 = m_3 - m_2 = 0$, $Q(m_1, m_2, m_3) = m_1$.

We now derive a closed-form expression for $Q(0, 1, 1 + d) = (q_1(d), q_2(d), q_3(d))$, say. By reversal equivariance, $q_3(d) = 1 + d - dq_1(\frac{1}{d})$ if $d \neq 0$. If $d = 0$, the median-interpolant of $(0, 1, 1 + d)$, π, is given by [L] of Proposition 2.2, with its maximum $x^* = 2$. A simple calculation gives $q_3(0) = \pi(\frac{11}{6}) = \frac{10}{9}$. We now work on obtaining expressions for q_1 and q_2.

PROPOSITION 2.3 (median-refinement, $D = 2$).

$$q_1(d) = \begin{cases} \frac{59}{27} + \frac{7}{27}d - \frac{8}{27}\sqrt{16 + 16d + d^2} & \textit{if } d \in [\frac{7}{3}, 5], \\ \frac{26}{27} + \frac{16}{27}d - \frac{4}{27}\sqrt{1 + 16d + 16d^2} & \textit{if } d \in [\frac{1}{5}, \frac{3}{7}], \\ \frac{77}{135} + \frac{13}{135}d + \frac{8}{135}\sqrt{1 - 62d + d^2} & \textit{if } d \in [-3, -\frac{1}{3}], \\ -\frac{1}{288}\frac{323 - 214d + 35d^2}{-1 + d} & \textit{if } d \in [-11, -3], \\ \frac{7}{9} - \frac{d}{9} & \textit{otherwise}; \end{cases} \tag{2.25}$$

$$q_2(d) = \begin{cases} -\frac{1}{270}\frac{1097 - 1174d + 17d^2 + (278 - 8d)\sqrt{1 - 62d + d^2}}{-4 + 4d - \sqrt{1 - 62d + d^2}} & \textit{if } d \in [-\frac{10}{7}, -\frac{7}{10}], \\ \frac{23}{30} + \frac{7}{30}d + \frac{1}{15}\sqrt{1 - 62d + d^2} & \textit{if } d \in [-3, -\frac{1}{3}] \backslash [-\frac{10}{7}, -\frac{7}{10}], \\ 1 & \textit{otherwise}. \end{cases} \tag{2.26}$$

Proof. Let π denote the median-interpolant of $(0, 1, 1 + d)$ associated with intervals $[0, 1], [1, 2], [2, 3]$. Recall that median-interpolation follows four branches [N1]–[N3] and [L] (cf. Proposition 2.2), whereas median-imputation follows two branches [N] and [L] (cf. Proposition 2.1.) The main task is to identify ranges of d for which median-interpolation and median-imputation use specific combinations of branches. The refinement result can then be described by obtaining algebraic expressions for each of those ranges. The calculations are similar to those in the proof of Proposition 2.2.

For q_1, there are five distinct cases:

1. $d \in [\frac{7}{3}, 5] \Rightarrow x^* \in [\frac{1}{4}, \frac{3}{4}]$: interpolation by branch [N1] and imputation by branch [L];
2. $d \in [\frac{1}{5}, \frac{3}{7}] \Rightarrow x^* \in [\frac{9}{4}, \frac{11}{4}]$: interpolation by branch [N2] and imputation by branch [L];
3. $d \in [-3, -\frac{1}{3}] \Rightarrow x^* \in [\frac{5}{4}, \frac{7}{4}]$: interpolation by branch [N3] and imputation by branch [L];

4. $d \in [-11, -3] \Rightarrow x^* \in [\frac{13}{12}, \frac{5}{4}]$: interpolation by branch [L] and imputation by branch [N];
5. $d \notin [\frac{7}{3}, 5] \cup [\frac{1}{5}, \frac{3}{7}] \cup [-3, -\frac{1}{3}] \cup [-11, -3] \Rightarrow x^* \in [\frac{13}{12}, \frac{5}{4}]$: interpolation by branch [L] and imputation by branch [L].

In each case, use the corresponding formulae in Proposition 2.2 to calculate π and then evaluate $q_1(d) = \mathrm{med}(\pi|[1, \frac{4}{3}])$ by Proposition 2.1.

For q_2, there are three distinct cases:

1. $d \in [-\frac{10}{7}, -\frac{7}{10}] \Rightarrow x^* \in [\frac{17}{12}, \frac{19}{12}]$: interpolation by branch [N3] and imputation by branch [N];[1]
2. $d \in [-3, -\frac{10}{7}] \cup [-\frac{7}{10}, -\frac{1}{3}] \Rightarrow x^* \in [\frac{5}{4}, \frac{17}{12}] \cup [\frac{19}{12}, \frac{7}{4}]$: interpolation by branch [N3] and imputation by branch [L];
3. $d \notin [-3, -\frac{1}{3}] \Rightarrow x^* \notin [\frac{5}{4}, \frac{7}{4}]$: $\mathrm{med}(\pi|[1, 2]) = \mathrm{med}(\pi|[4/3, 5/3]) = \pi(3/2) \equiv 1$.

In the first two cases, again use the corresponding formulae in Proposition 2.2 to calculate π followed by evaluating $q_2(d) = \mathrm{med}(\pi|[\frac{4}{3}, \frac{5}{3}])$ using Proposition 2.1. □

2.1.4. $D > 2$. Higher degree median-interpolation is also well-posed: [Q1] in section 2.1 has an affirmative answer for *all* integers A. A nonconstructive proof was found independently by the second author and Goodman [18]. However, it seems difficult to obtain a closed-form expression for the nonlinear refinement operator in case $D > 2$. It is possible to develop an iterative algorithm for MI that seems to converge exponentially fast to the median-interpolating polynomial for orders $D > 2$; see [15]. Experience with this algorithm suggests that MI is stable even for orders $D > 2$.

2.2. Multiscale refinement. The two-scale refinement scheme described in section 2 applied to an initial median sequence $(\tilde{m}_{j_0,k})_k \equiv (m_{j_0,k})_k$ implicitly defines a (generally nonlinear) refinement operator $R_{MI} = R$

$$R((\tilde{m}_{j,k})_k) = (\tilde{m}_{j+1,k})_k, \qquad j \geq j_0. \tag{2.27}$$

We can associate resulting sequences $(\tilde{m}_{j,k})_k$ with piecewise constant functions on the line via

$$\tilde{f}_j(\cdot) = \sum_{k=-\infty}^{\infty} \tilde{m}_{j,k}\, 1_{I_{j,k}}(\cdot) \text{ for } j \geq j_0. \tag{2.28}$$

This defines a sequence of piecewise constant functions defined on successively finer and finer meshes.

In case $D = 0$, we have

$$\tilde{f}_{j_0+h} = f_{j_0} \text{ for all } h \geq 0,$$

so the result is just a piecewise constant object taking value $m_{j_0,k}$ on $I_{j_0,k}$.

In case $D = 2$, we have no closed-form expression for the result. The operator R is nonlinear, and proving the existence of a limit $\tilde{f}_{j+h}$ as $h \to \infty$ requires work.

We mention briefly what can be inferred about multiscale average-interpolation from experience in subdivision schemes and in wavelet analysis. Fix $D \in \{2, 4, \dots\}$, and let $\overline{R} = \overline{R}^{(D)}$ denote the average-interpolation operator implicitly defined by (2.6).

[1] It is worth mentioning that this is the only case where *both* the interpolation and imputation are done using *nonlinear* rules.

Set $a_{0,k} = 1_{\{k=0\}}$. Iteratively refine this sequence by the rule $(a_{j+1,k})_k = \overline{R}((a_{j,k})_k)$. Define

$$\overline{f}_j(\cdot) = \sum_{k=-\infty}^{\infty} a_{j,k}\, 1_{I_{j,k}}(\cdot) \qquad \text{for } j \geq j_0. \tag{2.29}$$

The resulting sequence of $\overline{f}_j$ converges as $j \to \infty$ to a continuous limit $\phi = \phi^{(D)}$. This is called the fundamental solution of the multiscale refinement process. Due to the linearity of average-interpolation, if we refine an arbitrary bounded sequence $(a_{j_0,k})_k$ we get a continuous limit which is a superposition of shifted and dilated fundamental solutions:

$$\overline{f}(t) = \sum a_{j_0,k}\phi(2^{j_0}t - k). \tag{2.30}$$

For median-interpolation, such a superposition result cannot hold because of the nonlinearity of the refinement scheme for $D = 2, 4, \ldots$.

Figure 2.3 illustrates the application of multiscale refinement. Panel (a) shows the $D = 2$ refinement of three Kronecker sequences $m_{0,k}^{k'} = 1_{\{k=k'\}}$, $k' = 0, 1, 2$, as well as refinement of a Heaviside sequence $1_{\{k \geq 3\}}$. Panel (b) shows the $D = 2$ refinement of a Heaviside sequence $1_{\{k \geq 0\}}$. The sequence refined in (b) is the superposition of sequences refined in (a). Panel (c) gives a superposition of shifts of (a) for $k \geq 0$; if an analogue of (2.30) held for median refinement, this should be equal to panel (b). Panel (d) gives the discrepancy, (b)–(c). Note the vertical scales. While the discrepancy from "superposability" is not large, it is definitely nonzero and not simply an artifact of rounding or other numerical processes.

3. Convergence of median-interpolation, $D = 2$.

We now study some convergence properties of iterative median-interpolation. It turns out that for any bounded sequence $(m_{0,k})_k$, the sequence of nonlinear refinements $\tilde{f}_j$ converges to a bounded uniformly continuous limit $f(t)$. Moreover the limit has global Hölder exponent $\alpha > 0$. In this section, we will simplify notation and "drop tildes"; we denote a typical member of a refinement sequence by $m_{j,k}$ rather than $\tilde{m}_{j,k}$.

3.1. Weak convergence and stability.

Let Q be the refinement operator as defined in (2.23), and denote $Q^j = Q \circ \cdots \circ Q$ (Q composed with itself j times). We first show that, with any initial sequence $\{m_{j_0,k}\}$, $\{f_j\}$ converges at a dense set of points.

LEMMA 3.1. *For any* $[m_1, m_2, m_3] \in \mathbb{R}^3$, *the limit* $\lim_{j\to\infty} Q^j([m_1, m_2, m_3])$ *exists.*

Proof. Let $T_{j,k}$ denote the triple of intervals $[I_{j,k-1}, I_{j,k}, I_{j,k+1}]$. If π is the median-interpolant of $[m_1, m_2, m_3]$ on $T_{j_0,k}$, then it is also the interpolant for $Q([m_1, m_2, m_3])$ on the triple $T_{j_0+1,3k+1}$ arising from triadic subdivision of the central interval of $T_{j_0,k}$. If we refine the central subinterval of $T_{j_0+1,3k+1}$, we see that π continues to be the interpolant of the resulting medians. In fact, π is the interpolant for $Q^j([m_1, m_2, m_3])$ on the triple arising from the jth such generation of subdivision of central intervals, i.e., for $T_{j_0+j,\, 3^jk+(3^j-1)/2}$ for every $j > 0$. As $j \to \infty$, the sequence of sets $(T_{j_0+j,\, 3^jk+(3^j-1)/2})_j$ collapses to the midpoint of $I_{j_0,k}$. Therefore, by continuity of π and continuity of the imputation operator,

$$\lim_{j\to\infty} Q^j([m_1, m_2, m_3]) = [m, m, m], \text{ where } m = \pi\left(3^{-j_0}\left(k + \frac{1}{2}\right)\right),$$

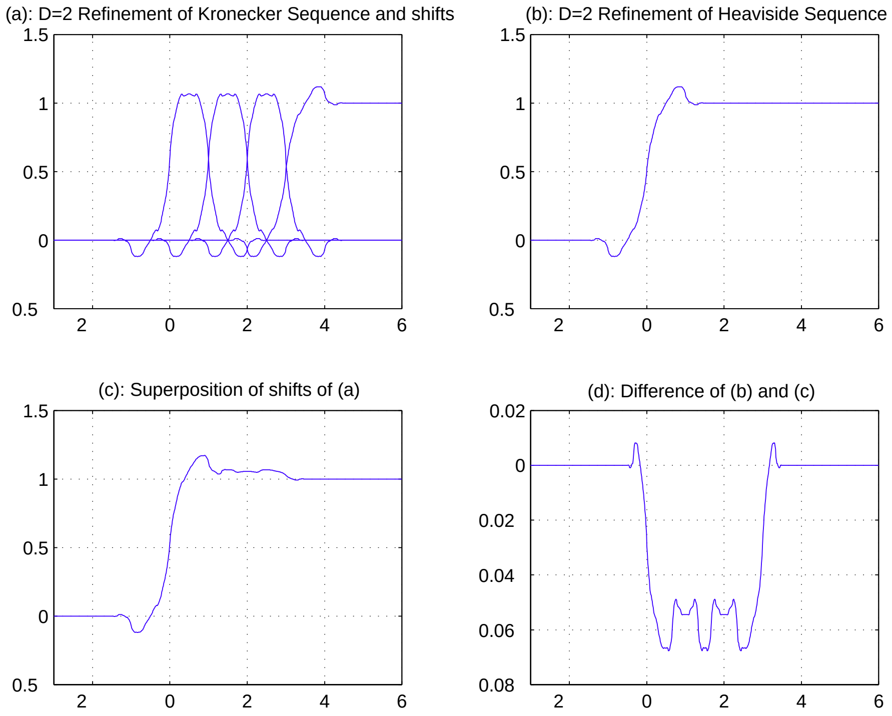


FIG. 2.3. *Discrepancy from "superposability" of multiscale median-interpolating refinement.*

the value of π at the midpoint of $I_{j_0,k}$. □

LEMMA 3.2 (convergence at triadic rationals). *For any initial median sequence* $\{m_{j_0,k}\}$*, the (nonlinear) iterative refinement scheme based on quadratic median-interpolation converges on a countable dense subset of the real line, i.e., there exists a countable dense set* $S \subset \mathbb{R}$ *and a function* $f : S \to \mathbb{R}$ *such that* $\lim_{j\to\infty} f_j(x) = f(x)$ *for every* $x \in S$.

Proof. Let $t_{j,k}$ be the midpoint of the triadic interval $I_{j,k}$. Assume we have applied the refinement scheme j_1 times to the input sequence $\{m_{j_0,k}\}_k$ (so that the values $m_{j_0+j_1,k-1}$, $m_{j_0+j_1,k}$, and $m_{j_0+j_1,k+1}$ have been calculated). We then have, for every $j > 0$, $f_{j_0+j_1+j}(t_{j_0+j_1,k}) = m^{(j)}$, where $m^{(j)}$ is the middle entry of $Q^j([m_{j_0+j_1,k-1}, m_{j_0+j_1,k}, m_{j_0+j_1,k+1}])$. By Lemma 3.1, $(m^{(j)})_j$ converges to a definite value as $j \to \infty$. We may take $S = \{t_{j,k} \mid j \geq j_0, k \in \mathbb{Z}\}$, the set of midpoints of all arbitrarily small triadic intervals, which is dense in $\mathbb{R}$. □

LEMMA 3.3. *For any* $j > 0$ *and* $k \in \mathbb{Z}$*, let* $d_{j,k} = m_{j,k} - m_{j,k-1}$*. Then*

$$|f(t_{j,k}) - m_{j,k}| \leq \frac{1}{15} \max\{|d_{j,k-1}|, |d_{j,k}|\}, \tag{3.1}$$

where the upper bound is attained if and only if $d_{j,k-1} = -d_{j,k}$.

Proof. From the proof of Lemma 3.2, $f(t_{j,k}) = \pi(t_{j,k})$, where π median-interpolates $m_{j,k-1}$, $m_{j,k}$, and $m_{j,k+1}$ for the triple $[I_{j,k-1}, I_{j,k}, I_{j,k+1}]$. Thus $|f(t_{j,k}) - m_{j,k}| = |\pi(t_{j,k}) - m_{j,k}|$. Unless $d = d_{j,k+1}/d_{j,k} \in [-3, -1/3]$, $\pi(t_{j,k}) = m_{j,k}$ and (3.1) holds trivially. When $d = d_{j,k+1}/d_{j,k} \in [-3, -1/3]$, $\pi(x)$ is given by [N3] of Proposition 2.2.

Without loss of generality, we can work with $j = 0$, $k = 0$ and denote (for simplicity) $m_i = m_{0,i}$, $i = -1, 0, 1$. Without loss of generality, we assume $\max\{|m_0 - m_{-1}|, |m_1 - m_0|\} = 1$, $m_{-1} = 0$, $m_0 = 1$ and $d = \frac{m_1 - m_0}{m_0 - m_{-1}} = m_1 - 1 \in [-1, -\frac{1}{3}]$. Then $|\pi(t_{j,k}) - m_{j,k}|/\max(|d_1|, |d_2|) = |\pi(3/2) - 1|$, where $\pi(x)$ is given by [N3] in Proposition 2.2, and we have

$$\max \frac{|\pi(t_{j,k}) - m_{j,k}|}{\max(|d_1|, |d_2|)} = \max_{d \in [-1, -\frac{1}{3}]} |a + b(3/2) + c(3/2)^2 - 1|$$
$$= \max_{d \in [-1, -\frac{1}{3}]} \left| \frac{7}{30}(d - 1) + \frac{\sqrt{1 - 62d + d^2}}{15} \right| = \frac{1}{15}.$$

The maximum is attained at $d = -1$. □

3.2. Hölder continuity. We now develop a basic tool for establishing Hölder continuity of refinement schemes.

Theorem 3.4. *Let $(m_{0,k})_k$ be a bounded sequence, and $m_{j,k}$ be the refinement sequences generated by the quadratic median-interpolating refinement scheme constructed in section 2.2. Let $d_{j,k} := m_{j,k} - m_{j,k-1}$ and $f_j := \sum_k m_{j,k} 1_{I_{j,k}}$. Suppose that for some $\alpha > 0$ and all $j \geq 0$, $\sup_k |d_{j,k}| \leq C \cdot 3^{-j\alpha}$. Then f_j converges uniformly to a bounded limit $f \in \dot{C}^\alpha$. The converse is also true if $\alpha \leq 1$.*

This is analogous to results found in the literature of linear refinement schemes (c.f. Theorem 8.1 of Rioul [25]). The proof of the forward direction uses basically the same arguments as in the linear case, except that one must deal with nonlinearity using a general affine-invariance property of medians. Similar arguments could be applied in the study of cases $D > 2$. The proof of the converse direction, on the other hand, relies on Lemma 3.3 and is therefore specific to the $D = 2$ triadic case.

Proof. ($\Rightarrow$) We show that $\{f_j\}$ is a Cauchy sequence. Consider

$$\sup_x |f_{j+1}(x) - f_j(x)| = \sup_k \max_{\epsilon = 0,1,2} |m_{j+1,3k+\epsilon} - m_{j,k}|. \tag{3.2}$$

The functions $m_{j+1,3k+\epsilon} = q_\epsilon(m_{j,k-1}, m_{j,k}, m_{j,k+1})$ obey

$$q_\epsilon(m_{j,k-1}, m_{j,k}, m_{j,k+1}) = m_{j,k} + q_\epsilon(-d_{j,k}, 0, d_{j,k+1}).$$

Moreover, by Lemma 6.2, these functions are Lipschitz: $q_\epsilon(m_1, m_2, m_3) \leq c \max_{i=1,2,3}\{|m_i|\}$. Therefore,

$$\sup_x |f_{j+1}(x) - f_j(x)| = \sup_k \max_{\epsilon = 0,1,2} |q_\epsilon(-d_{j,k}, 0, d_{j,k+1})| \leq c \sup_k |d_{j,k}| \leq c \cdot C \cdot 3^{-j\alpha}$$

and $||f_{j+p} - f_j||_\infty \leq cC(3^{-j\alpha} + \cdots + 3^{-(j+p-1)\alpha}) \leq C' 3^{-j\alpha}$. ($C'$ is independent of p.) Hence $\{f_j\}$ is a Cauchy sequence that converges uniformly to a function f. Furthermore, $f \in \dot{C}^\alpha$ because $\sup_{3^{-(j+1)} \leq |h| \leq 3^{-j}} |f(x+h) - f(x)| \leq |f(x+h) - f_j(x+h)| + |f(x) - f_j(x)| + |f_j(x+h) - f_j(x)| \leq C' 3^{-j\alpha} + C' 3^{-j\alpha} + C 3^{-j\alpha} \leq 3^\alpha (2C' + C)|h|^\alpha$.

($\Leftarrow$) If $f \in \dot{C}^\alpha$, $\alpha \leq 1$, then, by definition, $\sup_k |f(t_{j,k}) - f(t_{j,k-1})| \leq c \cdot 3^{-j\alpha}$. But

$$|m_{j,k} - m_{j,k-1}| \leq |m_{j,k} - f(x_{j,k})| + |f(t_{j,k}) - f(t_{j,k-1})| + |m_{j,k-1} - f(t_{j,k-1})|$$
$$\leq \frac{1}{15} \max\{|d_{j,k}|, |d_{j,k-1}|\} + c \cdot 3^{-j\alpha} + \frac{1}{15} \max\{|d_{j,k-1}|, |d_{j,k-2}|\}.$$

The last inequality is due to Lemma 3.3. Maximizing over k on both sides of the above inequality, followed by collecting terms, gives $\sup_k |d_{j,k}| \leq (15/13\ c) \cdot 3^{-j\alpha}$. □

3.3. Nonlinear difference scheme. As in Theorem 3.4, let $d_{j,k} = m_{j,k} - m_{j,k-1}$ denote the sequence of interblock differences. It is a typical property of any *constant-reproducing* linear refinement scheme that the difference sequences can themselves be obtained from a linear refinement scheme, called the *difference scheme*. The coefficient mask of that scheme is easily derivable from that of the original scheme; see [16, 25]. More generally, a linear refinement scheme that can reproduce all lth degree polynomials would possess $l+1$ difference (and divided difference) schemes [16]. A partial analogy to this property holds in the nonlinear case: the $D = 2$ median-interpolation scheme, being a nonlinear refinement scheme with quadratic polynomial reproducibility, happens to possess a (nonlinear) first difference scheme but no higher order ones.

Let $d_{j,k} \neq 0, d_{j,k+1} \neq 0, d_{j,k+2}$ be given. Then, by (2.24),

$$\begin{aligned}
(m_{j+1,3k}, &m_{j+1,3k+1}, m_{j+1,3k+2}) \\
&= m_{j,k-1} + d_{j,k}\, Q\left(0, 1, 1 + \frac{d_{j,k+1}}{d_{j,k}}\right) \\
&= m_{j,k-1} + d_{j,k}\, \left(q_1\left(\frac{d_{j,k+1}}{d_{j,k}}\right), q_2\left(\frac{d_{j,k+1}}{d_{j,k}}\right), q_3\left(\frac{d_{j,k+1}}{d_{j,k}}\right)\right), \\
(m_{j+1,3k+3}, &m_{j+1,3k+4}, m_{j+1,3k+5}) \\
&= m_{j,k} + d_{j,k+1} Q\left(0, 1, 1 + \frac{d_{j,k+2}}{d_{j,k+1}}\right) \\
&= m_{j,k-1} + d_{j,k} + d_{j,k+1}\left(q_1\left(\frac{d_{j,k+2}}{d_{j,k+1}}\right), q_2\left(\frac{d_{j,k+2}}{d_{j,k+1}}\right), q_3\left(\frac{d_{j,k+2}}{d_{j,k+1}}\right)\right).
\end{aligned}$$

Hence $d_{j+1,3k+1} d_{j+1,3k+2}, d_{j+1,3k+3}$ are only dependent on $d_{j,k}, d_{j,k+1}, d_{j,k+2}$ and there exist three functionals $\partial q_0 : \mathbb{R}^2 \to \mathbb{R}$, $\partial q_1 : \mathbb{R}^2 \to \mathbb{R}$, $\partial q_2 : \mathbb{R}^3 \to \mathbb{R}$ such that $d_{j+1,3k+1} = \partial q_0(d_{j,k}, d_{j,k+1})$, $d_{j+1,3k+2} = \partial q_1(d_{j,k}, d_{j,k+1})$, and $d_{j+1,3k+3} = \partial q_2(d_{j,k}, d_{j,k+1}, d_{j,k+2})$, where

$$\begin{aligned}
\partial q_0(d_0, d_1) &= d_0\left(q_2\left(\frac{d_1}{d_0}\right) - q_1\left(\frac{d_1}{d_0}\right)\right) \quad \text{(when } d_0 \neq 0\text{)}, \\
\partial q_1(d_0, d_1) &= d_0\left(q_3\left(\frac{d_1}{d_0}\right) - q_2\left(\frac{d_1}{d_0}\right)\right) \quad \text{(when } d_0 \neq 0\text{)}, \\
\partial q_2(d_0, d_1, d_2) &= d_0 + d_1 q_1\left(\frac{d_2}{d_1}\right) - d_0 q_3\left(\frac{d_1}{d_0}\right) \quad \text{(when } d_0 \neq 0 \text{ and } d_1 \neq 0\text{)}, \\
&= d_0 + d_1 q_1\left(\frac{d_2}{d_1}\right) - d_0\left(1 + \frac{d_1}{d_0} - \frac{d_1}{d_0} q_1\left(\frac{d_0}{d_1}\right)\right). \qquad (3.3)
\end{aligned}$$

The degenerate cases can be handled easily. One of those will be of use later, namely,

$$\partial q_0(0, d_1) = \frac{d_1}{9} = \lim_{d_0 \to 0} d_0\left(q_2\left(\frac{d_1}{d_0}\right) - q_1\left(\frac{d_1}{d_0}\right)\right). \qquad (3.4)$$

Similar limits hold for ∂q_1 and ∂q_2.

The difference scheme inherits two nice equivariance properties from median-interpolation:

- *Reversal equivariance.*

$$\begin{aligned}
\partial q_0(d_1, d_0) = \partial q_1(d_0, d_1), &\partial q_1(d_1, d_0) \\
&= \partial q_0(d_0, d_1), \text{ and } \partial q_2(d_2, d_1, d_0) = \partial q_2(d_0, d_1, d_2).
\end{aligned}$$

- *Affine equivariance.*

$$\partial q_\epsilon(b(d_0,d_1)) = b\,\partial q_\epsilon(d_0,d_1), \epsilon = 0,1, \text{ and}$$
$$\partial q_2(b(d_0,d_1,d_2)) = b\,\partial q_2(d_0,d_1,d_2).$$

The above discussion implies the existence of three (nonlinear) operators $\partial Q_\epsilon : \mathbb{R}^3 \to \mathbb{R}^3$, $\epsilon = 0,1,2$ that govern the difference scheme:

$$\partial Q_\epsilon([d_{j,k-1}, d_{j,k}, d_{j,k+1}]^T) = [d_{j+1,3k+\epsilon}, d_{j,3k+\epsilon}, d_{j,3k+\epsilon}]^T \text{ for all } j \geq 0, k \in \mathbb{Z}. \tag{3.5}$$

Uniform convergence will follow from the fact that these operators are *shrinking* in the sense that

$$S_\infty(\partial Q_\epsilon) := \max_{d\in\mathbb{R}^3} \frac{||\partial Q_\epsilon(d)||_\infty}{||d||_\infty} = \beta < 1, \quad \epsilon = 0,1,2. \tag{3.6}$$

As the ∂Q_ϵ are nonlinear, this is slightly weaker than being *contractive.* We will prove an inequality like this in the next section.

It is easy to check that $\partial Q_\epsilon(d) = 0$ if and only if $d = 0$ and that $S_\infty(\partial Q_{\epsilon_1} \circ \partial Q_{\epsilon_2}) \leq S_\infty(\partial Q_{\epsilon_1}) S_\infty(\partial Q_{\epsilon_2})$. In order to bound the decay rate of $\max_k |d_{j,k}|$ (and hence the critical Hölder exponent for median-interpolating refinements), we can use the estimate

$$\sup_k |d_{j,k}| \leq \sup_k |d_{0,k}| \max_{\epsilon_i=0,1,2} S_\infty(\partial Q_{\epsilon_j} \circ \cdots \circ \partial Q_{\epsilon_1}). \tag{3.7}$$

Assuming (3.6), we can bound the right-hand side of (3.7) crudely by

$$\max_{\epsilon_i=0,1,2} S_\infty(\partial Q_{\epsilon_j} \circ \cdots \circ \partial Q_{\epsilon_1}) \leq \max_{\epsilon_i=0,1,2} S_\infty(\partial Q_{\epsilon_j}) \times \cdots \times S_\infty(\partial Q_{\epsilon_1}) \tag{3.8}$$
$$= \beta^j = 3^{-j\,\alpha}, \tag{3.9}$$

where $\alpha = \log_3(1/\beta) > 0$. Hence, uniform convergence follows from Theorem 3.4.

Actually, the inequality (3.8) contains slack. It is possible to improve on it by adapting to the nonlinear case approaches developed by Rioul [25] and Dyn, Gregory, and Levin [16] in the study of linear refinement schemes. We state without proof the following: Define α_j by

$$3^{-j\alpha_j} = \max_{\epsilon_i=0,1,2} S_\infty(\partial Q_{\epsilon_j} \circ \cdots \circ \partial Q_{\epsilon_1}). \tag{3.10}$$

Let $\alpha := \sup_j \alpha_j$. Then $\lim_j \alpha_j = \alpha$ and median-interpolating refinements are $\dot{C}^{\alpha-\epsilon}$ for $\epsilon > 0$. This observation is potentially useful because it provides a way to compute *lower bounds* for the Hölder regularity of median-interpolating refinement limits. In the next section, we apply this idea with the choice of $j = 1$, which results in the crude lower bound $\alpha_1 = \log_3(135/121)$. A better bound might be obtained if one could manage to compute the right-hand side of (3.10) for a larger j.

3.4. The difference scheme is shrinking. Armed with the closed-form expression for the quadratic median-interpolating refinement scheme, we can explicitly calculate $S_\infty(\partial Q_\epsilon)$ despite the nonlinearity of the operator.

THEOREM 3.5. *$S_\infty(\partial Q_\epsilon) = 121/135 < 1$ for $\epsilon = 0,1,2$. Consequently, by Theorem* 3.4 *and* (3.7)–(3.9)*, for any bounded initial sequence $m_{0,k}$, the sequence of* *nonlinear refinements* $f_j = \sum_k m_{j,k} 1_{I_{j,k}}$ *converges uniformly to a bounded uniformly continuous function* $f \in \dot{C}^\alpha$*, where* $\alpha = \log_3(135/121) \approx 0.0997$.

Proof. See [31] for the computational details. The main idea of the proof is to verify that $\partial q_0(d_0, d_1)$ and $\partial q_1(d_0, d_1)$ are monotone increasing in d_0 for fixed d_1 and monotone increasing in d_1 for fixed d_0; and that $\partial q_2(d_0, d_1, d_2)$ is monotone decreasing in d_0 and d_2 for fixed d_1 and monotone increasing in d_1 for fixed d_0 and d_2. Thus $S_\infty(\partial q_0) := \max_{|d_0|,|d_1| \le 1} \partial q_0(d_0, d_1) = \partial q_0(1, 1) = 1/3$, and, by symmetry, $S_\infty(\partial q_1) = S_\infty(\partial q_0) = 1/3$; $S_\infty(\partial q_2) = \partial q_2(-1, 1, -1) = 121/135$. The theorem follows from the fact that $S_\infty(\partial Q_\epsilon) = \max_{i=0,1,2} S_\infty(\partial q_i)$ for $\epsilon = 0, 1, 2$. □

3.5. Discussion.

The regularity bound $\alpha \ge \log_3(135/121) \approx 0.0997$ is probably very far from sharp. We now discuss evidence suggesting that the sharp Hölder exponent is nearly 1.

3.5.1. Linearized median-interpolation.

We recall from Figure 2.2, and Propositions 2.1, 2.2, and 2.3 that there is an underlying *linear branch* associated with the median scheme. A sufficient but not necessary condition for applicability of this branch is that the block medians be consistent with a polynomial π that is monotone throughout $[a, b]$.

In the linear branch, the median functional amounts to *midpoint evaluation*: $\mathrm{med}(\pi|[a, b]) = \pi((a + b)/2)$. The resulting refinement rule is a linear scheme that we call the *LMI* scheme, with coefficient mask $[-1/9, 0, 2/9, 7/9, 1, 7/9, 2/9, 0, -1/9]$. It is a symmetric interpolatory scheme and can be viewed as a triadic variant of Deslauriers–Dubuc schemes. The mask has a positive Fourier transform, and the convergence and critical Hölder regularity of the scheme can be determined quite easily by applying the theory of linear refinement schemes [25, 5, 8].

The LMI scheme has refinement limits which are "almost Lipschitz" [31]. For any given bounded initial sequence of block values at scale 0, the LMI scheme converges to a bounded uniformly continuous limit f obeying the regularity estimate $\sup_x |f(x + h) - f(x)| \le C|h| \log(1/|h|)$. Moreover, the above global regularity bound cannot be improved. (If we study *local* rather than global regularity, it can be shown, using techniques in [5], that the bound can be improved for *most* x.)

See Figure 3.1 for pictures of median-interpolating and linearized median-interpolating refinement limits of the Kronecker sequence $\{m_{0,k} = \delta_{0,k}\}$.

3.5.2. Critical Hölder exponent conjectures.

We conjecture that MI and LMI share the same global Hölder regularity. This is a rather natural conjecture to make, since the difference between MI and LMI is actually very small—as one sees from the near-linearity of the functions displayed in Figure 2.2. In [31], computational evidence was provided to support the conjecture. In particular, the experiments there suggest the following:

1. The actual decay behavior in (3.10) is $O(j3^{-j})$, which is much faster than the rate bound calculated in Theorem 3.5. This rate would imply that median-interpolating refinement limits are almost Lipschitz. See panel (d) of Figure 3.1.
2. Both the MI refinement sequences $m^{MI}_{j,k}$ and the LMI refinement sequences $m^{LMI}_{j,k}$ appear to possess a *stationarity property*: let k^*_j be the value of k maximizing $|d^{MI}_{j,k}|$; here there exists an integer k^* such that $3^{-j} k^*_j = k^*$ for all large enough j. See panel (b) of Figure 3.1. The same phenomenon is observed for $d^{LMI}_{j,k}$; see panel (c) of Figure 3.1. Stationarity is a *provable*

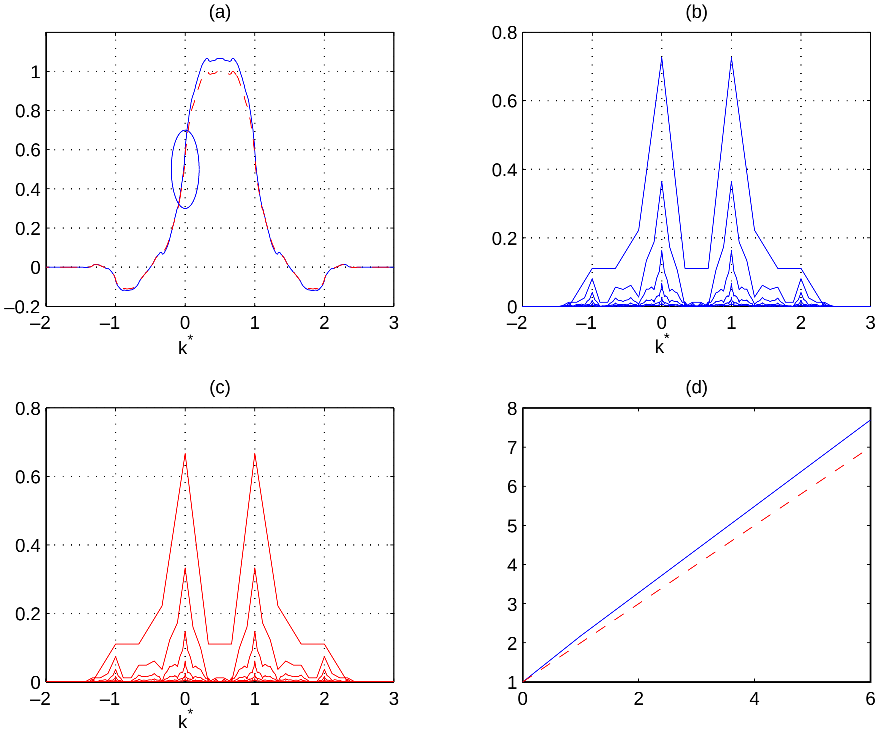


Fig. 3.1. *MI versus LMI:* (a) *MI- and LMI- refinements (solid and dashed lines, respectively) of* $m_{0,k} = \delta_{0,k}$, (b) $|d_{j,k}^{MI}|$ *versus* $k3^{-j}$, $j = 1, \ldots, 6$, (c) $|d_{j,k}^{LMI}|$ *versus* $k3^{-j}$, $j = 1, \ldots, 6$, (d) $3^j \max_k |d_{j,k}^{MI}|$ *and* $3^j \max_k |d_{j,k}^{LMI}|$ *versus* j *(solid and dashed lines, respectively).*

property of the LMI scheme. It is *empirically* a property of the MI scheme as well.

3. It appears that in the vicinity of the spatial location $x = k^*$, the limit function is monotone and (consequently) median-interpolating refinement is repeatedly using its *linear branch*. Therefore, it seems that $\sup_k |d_{j,k}^{MI}|$ and $\sup_k |d_{j,k}^{LMI}|$ share the same asymptotics. See again Figure 3.1.

A more ambitious open question is the following: Let $x \in \mathbb{R}$, and let $k_j(x)$ be defined by $x \in I_{j,k_j(x)}$. We call x an *asymptotically linear point* if, for large enough j, median-interpolation is only using its linear branch to determine $m_{j+1,k_{j+1}(x)}$ from $m_{j,k_j(x)+\epsilon}$, $\epsilon = -1, 0, 1$. In order to understand deeply the relation between median-interpolation and linearized median-interpolation, it would be useful to determine the structure of the set of asymptotically linear points.

4. Median-interpolating pyramid transform. We now apply the refinement scheme to construct a nonlinear pyramid and associated nonlinear multiresolution analysis.

4.1. Pyramid algorithms. While it is equally possible to construct pyramids for decomposition of functions $f(t)$ or of sequence data y_i, we keep an eye on applications and concentrate attention on the sequence case. So we assume we are given a discrete dataset $y_i, i = 0, \ldots, n-1$, where $n = 3^J$ is a triadic number. We aim to use the nonlinear refinement scheme to decompose and reconstruct such sequences.

ALGORITHM FMIPT: PYRAMID DECOMPOSITION.

1. *Initialization.* Fix $D \in 0, 2, 4, \ldots$ and $j_0 \geq 0$. Set $j = J$.
2. *Formation of block medians.* Calculate
$$m_{j,k} = \mathrm{med}(y_i : i/n \in I_{j,k}). \tag{4.1}$$
(Here med() refers to the discrete median rather than the continuous median.)
3. *Formation of refinements.* Calculate
$$\tilde{m}_{j,k} = R((m_{j-1,k})) \tag{4.2}$$
using refinement operators of the previous section.
4. *Formation of detail corrections.* Calculate
$$\alpha_{j,k} = m_{j,k} - \tilde{m}_{j,k}.$$
5. *Iteration.* If $j = j_0 + 1$, set $m_{j_0,k} = \mathrm{med}(y_i : i/n \in I_{j_0,k})$ and terminate the algorithm, else set $j = j - 1$ and goto 2.

ALGORITHM IMIPT: PYRAMID RECONSTRUCTION.

1. *Initialization.* Set $j = j_0 + 1$. Fix $D \in 0, 2, 4, \ldots$ and $j_0 \geq 0$, as in the decomposition algorithm.
2. *Reconstruction by refinement.*
$$(m_{j,k}) = R((m_{j-1,k})) + (\alpha_{j,k})_k.$$
3. *Iteration.* If $j = J$ goto 4, else set $j = j + 1$ and goto 2.
4. *Termination.* Set
$$y_i = m_{J,i}, \quad i = 0, \ldots, n-1.$$

An implementation is described in [31]. Important details described there include the treatment of boundary effects and efficient calculation of block medians.

DEFINITION 4.1. *Gather the outputs of the pyramidal decomposition algorithm into the sequence*
$$\theta = ((m_{j_0,k})_k, (\alpha_{j_0+1,k})_k, (\alpha_{j_0+2,k})_k, \ldots, (\alpha_{J,k})_k).$$
We call θ the MIPT of y and we write $\theta = MIPT(y)$. Applying the pyramidal reconstruction algorithm to θ gives an array which we call the inverse transform, and we write $y = MIPT^{-1}(\theta)$.

The reader may wish to check that $MIPT^{-1}(MIPT(y)) = y$ for every sequence y.

We will also use below the average-interpolating pyramid transform (AIPT), defined in a completely parallel way, using only the average-interpolation refinement operator $\overline{R}$. We write $\theta = AIPT(y)$ and $y = AIPT^{-1}(\theta)$.

Complexity. Both transforms have good computational complexity. The refinement operator for $AIPT$, in common with wavelet transforms and other multiscale algorithms, has order $O(n)$ computational complexity. The coarsening operator can be implemented with the same complexity because of a *causality* relationship:
$$\mathrm{ave}(y_i|I_{j,k}) = \mathrm{ave}(\mathrm{ave}(y_i|I_{j+1,3k}),\ \mathrm{ave}(y_i|I_{j+1,3k+1}),\ \mathrm{ave}(y_i|I_{j+1,3k+2})). \tag{4.3}$$

Similarly, the refinement operator of $MIPT$ of order $D = 2$ has complexity $O(n)$ due to the propositions of section 2.1.3. However, for the coarsening operator there is no direct causality relationship. The analogue of (4.3) obtained by replacing "ave" by "med" does not hold.

To rapidly calculate all medians over triadic blocks, one can maintain sorted lists of the data in each triadic block; the key coarsening step requires merging three sorted lists to obtain a single sorted list. This process imposes only a $\log_3(n)$ factor in additional cost. For a more detailed description of the implementation, we refer to [31]. As a result, $MIPT$ can be implemented by an $O(n \log_3 n)$ algorithm, whereas $MIPT^{-1}$ can be implemented with $O(n)$ time-complexity.

4.2. Properties. P1. *Coefficient localization.* The coefficient $\alpha_{j,k}$ in the pyramid only depends on block medians of blocks at scale $j-1$ and j which cover or abut the interval $I_{j,k}$.

P2. *Expansionism.* There are 3^{j_0} résumé coefficients $(m_{j_0,k})$ in θ and 3^j coefficients $(\alpha_{j,k})_k$ at each level j. Hence

$$\mathrm{Dim}(\theta) = 3^{j_0} + 3^{j_0+1} + \cdots + 3^J.$$

It follows that $\mathrm{Dim}(\theta) = 3^J(1 + 1/3 + 1/9 + \cdots) \sim 3/2 \cdot n$. The transform is about 50% expansionist.

P3. *Coefficient decay.* Suppose that the data $y_i = f(i/n)$ are noiseless samples of a continuous function $f \in \dot{C}^\alpha$, $0 \le \alpha \le 1$, i.e., $|f(s) - f(t)| \le C|s-t|^\alpha$ for a fixed C. Then for $MIPT$ $D = 0$ or 2, we have

$$|\alpha_{j,k}| \le C'C3^{-j\alpha}. \tag{4.4}$$

Suppose f is $\dot{C}^{r+\alpha}$ for $r = 1$ or 2, i.e., $|f^{(r)}(s) - f^{(r)}(t)| \le C|s-t|^\alpha$, for some fixed α and C, $0 < \alpha \le 1$. Then, for $MIPT$ $D = 2$,

$$|\alpha_{j,k}| \le C'C3^{-j(r+\alpha)}. \tag{4.5}$$

P4. *Gaussian noise.* Suppose that $y_i = \sigma z_i$, $i = 0, \dots, n-1$, and that z_i is i.i.d $N(0,1)$, a standard Gaussian white noise. Then

$$P(\sqrt{3^{J-j}}\,|\alpha_{j,k}| \ge \xi) \le C_1 \cdot \exp\left(-C_2 \frac{\xi^2}{\sigma^2}\right),$$

where the $C_i > 0$ are absolute constants.

These properties are things we naturally expect of linear pyramid transforms, such as those of Adelson and Burt, and P1, P3, and P4 we expect also of wavelet transforms. In fact these properties hold not just for MIPT but also for AIPT.

A key property of MIPT but *not* AIPT is the following.

P5. *Cauchy noise.* Suppose that $y_i = \sigma z_i$, $i = 0, \dots, n-1$, and that z_i is i.i.d standard Cauchy white noise. Then

$$P(\sqrt{3^{J-j}}\,|\alpha_{j,k}| \ge \xi) \le C_1' \cdot \exp\left(-C_2' \frac{\xi^2}{\sigma^2}\right),$$

where $0 \le \xi \le \sqrt{3^{J-j}}$ and the $C_i' > 0$ are absolute constants.

For a linear transform, such as AIPT, the coefficients of Cauchy noise have Cauchy distributions, and such exponential bounds cannot hold. Moreover, the spread of the resulting Cauchy distributions does not decrease with increasing j. In contrast, P5 shows that the spread of the MIPT coefficients gets smaller with larger j, and that deviations more than a few multiples of the spread are very rare.

Properties P1 and P2 need no further proof; P3–P5 are proved in the appendix.

4.3. MRA. MI refinement allows us to mimic the multiresolution analysis of wavelet theory. Given the sequence $(m_{j,k})$ of block medians of y at scale j, we may apply $J-j$ iterations of two-scale refinement to these medians, getting a sequence of length n which we can call $P_j y$. An equivalent definition is as follows:

- Decomposition. $\theta = MIPT(y)$.
- Suppression of details. Let $\tilde{\theta}$ be a partial copy of θ, where we set $\alpha_{j',k} = 0$ for $j' > j$.
- Reconstruction. $P_j y = MIPT^{-1}(\tilde{\theta})$.

P_j is a nonlinear approximation of y at the scale j, because it uses only the block medians at scale j in its construction.

We can also form $Q_j y = P_j y - P_{j-1} y$, listing the details present in the approximation at scale j but not present at scale $j-1$.

4.4. Examples. We collect here a few examples of the MIPT for $D = 2$.

Figure 4.1 shows three different noiseless signals: (a) Sinusoid; (b) Heaviside; (c) Doppler. It also shows the pyramid coefficients of noiseless data for $D = 2$.

Figure 4.2 shows an MRA decomposition of the same three signals. This display shows $P_{j_0} y$, $Q_{j_0+1} y, \ldots, Q_J y$.

5. Denoising by MIPT thresholding. We now consider applications of pyramid transforms to multiscale denoising. In general, we act as we would in the wavelet denoising case.

- *Pyramid decomposition.* Calculate $\theta = MIPT(y)$.
- *Hard thresholding.* Let $\eta_t(y) = y \cdot 1_{\{|y|>t\}}$ be the hard thresholding function and let
$$\hat{\theta} = ((m_{j_0,k})_k, (\eta_{t_{j_0+1}}(\alpha_{j_0+1,k}))_k, \ldots).$$
Here the (t_j) is a sequence of threshold levels.
- *Pyramid reconstruction.* Calculate $\hat{f} = MIPT^{-1}(\hat{\theta})$.

In this approach, coefficient amplitudes smaller than t_j are judged negligible, as noise rather than signal. Hence the thresholds t_j control the degree of noise rejection but also of valid signal rejection. One hopes, in analogy with the orthogonal transform case studied in [12], to set thresholds which are small but which are very likely to exceed every coefficient in case of a pure noise signal. If the MIPT performs as we hope, the MIPT thresholds can be set "as if" the noise were Gaussian and the transform were AIPT, even when the noise is very non-Gaussian. This would mean that the median pyramid is immune to bad effects of impulsive noise.

5.1. Choice of thresholds. Motivated by P4 and P5, we work with the "L^2-normalized" coefficients $\bar{\alpha}_{j,k} = \sqrt{3^{J-j}}\alpha_{j,k}$ in this section.

In order to choose thresholds $\{t_j\}$ which are very likely to exceed every coefficient in case of a pure noise signal, we find t_j satisfying $P(|\bar{\alpha}_{j,k}| > t_j) \le c \cdot 3^{-J}/J$ where the MIPT coefficients arise from a pure noise signal $(X_i)_{i=0}^{3^J-1}$, $X_i \sim_{i.i.d.} F$. Then we have

$$P(\exists (j,k) \text{ s.t. } |\bar{\alpha}_{j,k}| > t_j) \le c \cdot \sum_{j=j_0}^{J} \sum_{k=0}^{3^j-1} \frac{1}{J} 3^{-J} \to 0 \text{ as } J \to \infty. \tag{5.1}$$

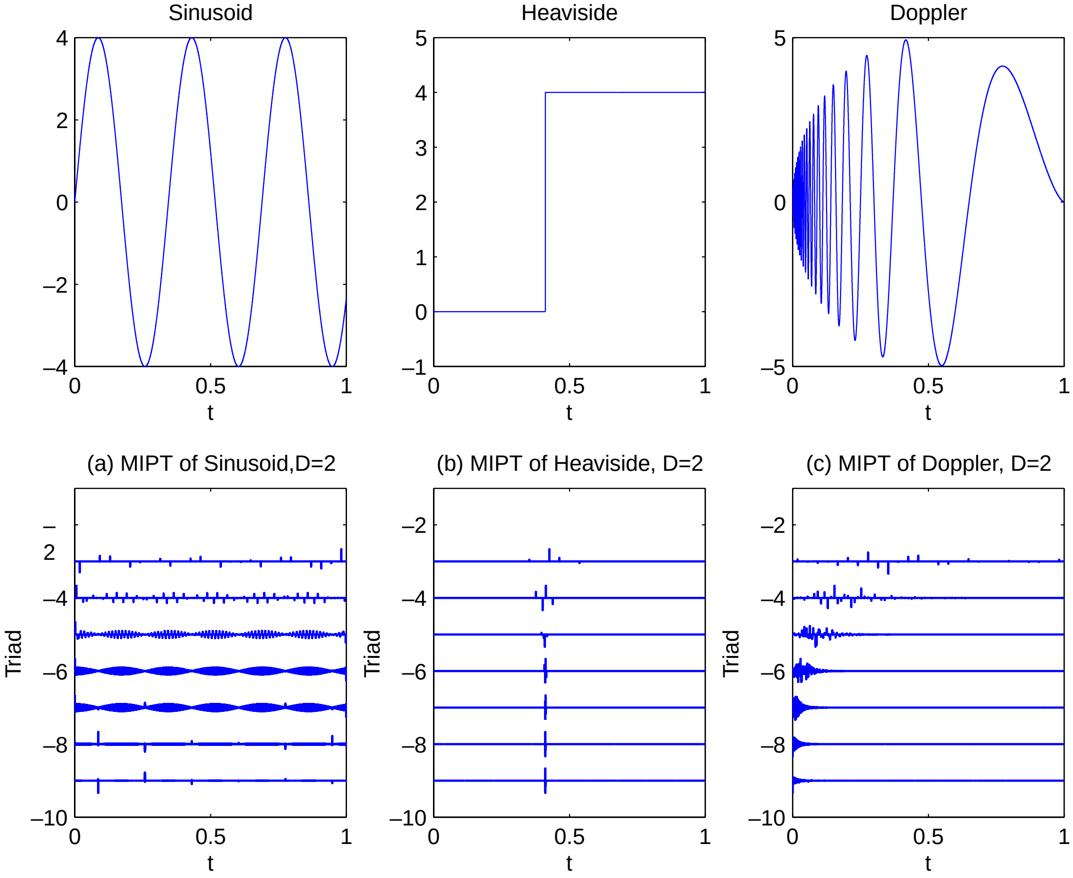


FIG. 4.1. *MIPT coefficients of three synthetic signals: Sinusoid, Heaviside, and Doppler. In each case, the plot of* $(\alpha_{j,k})_k$ *is scaled by* $\max_k |\alpha_{j,k}|$.

By (6.5), we can simply choose t_j satisfying $P(\sqrt{3^{J-j}}|\operatorname{med}(X_1, \dots, X_{3^{J-j}})| > t_j) \leq 3^{-J}/J$. Corollary 6.5 gives, when F is a symmetric law,

$$t_j := t_j(F) = \sqrt{3^{J-j}}F^{-1}\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \left(\frac{1}{2J3^J}\right)^{\frac{2}{3^{J-j}}}}\right). \tag{5.2}$$

Careful study of (5.2) suggests to us that away from the finest scales, the magnitude of t_j is governed by the behavior of F^{-1} near $1/2$. Hence after standardizing the level and slope of F at $p = 1/2$ we expect that the threshold depends very little on F.

The discussion of the last few paragraphs has been informal, but the "weak dependence of thresholds on F" can be formalized. Consider classes of smooth distributions $\mathcal{F}(M, \eta)$ defined as follows. First, the distributions have densities f symmetric about 0, so that $F^{-1}(1/2) = 0$. Second, scale is standardized so that each density obeys $f(0) = 1/\sqrt{2\pi}$, the same as the standard Gaussian $N(0, 1)$. This is of course equivalent to setting $(F^{-1})'(1/2) = \sqrt{2\pi}$. Third, we impose on $F^{-1}(p)$ some regularity near

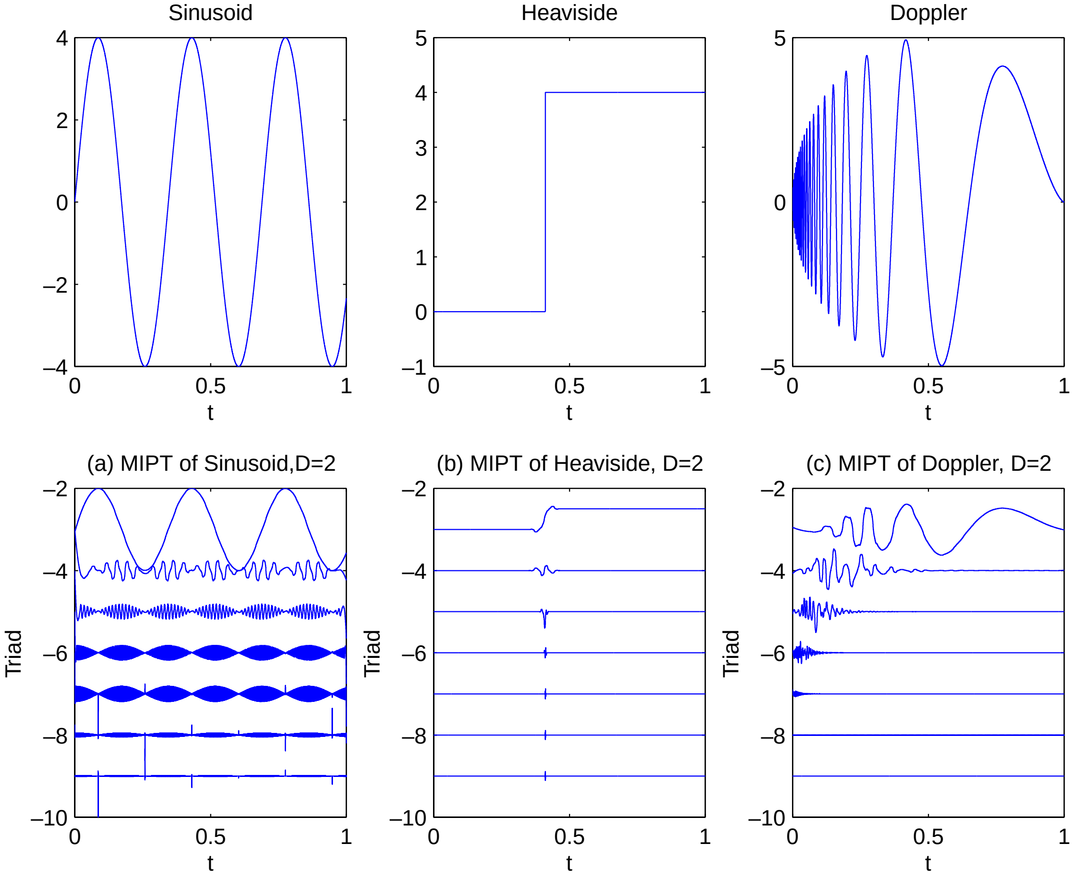


FIG. 4.2. *Nonlinear multiresolution analysis of three synthetic signals: Sinusoid, Heaviside, and Doppler. For Heaviside and Doppler, the plots of Q_j are scaled the same for all j; whereas for the Sinusoid, each Q_j is scaled by* $\max_k |(Q_j)_k|$.

$p = 1/2$: the existence of two continuous derivatives throughout $[1/2 - \eta, 1/2 + \eta]$. Our classes of symmetric distributions $\mathcal{F}(M, \eta)$ are then

$$\mathcal{F}(M,\eta) := \{F: \ f \text{ symmetric}, \ (F^{-1})'(1/2) = \sqrt{2\pi}, \ |(F^{-1})''(p)| \le M, \ |p - 1/2| \le \eta\},$$

where $M > 0$ and $0 < \eta < 1/2$ are absolute constants. The appendix proves the following theorem.

THEOREM 5.1. *For any $\epsilon > 0$ and $\theta \in (0,1)$, there exists $J^* = J^*(\epsilon, \theta, M, \eta)$ such that if $J \ge J^*$, then*

$$\max_{j \le \lfloor \theta J \rfloor} |t_j(F_1) - t_j(F_2)| \le \epsilon \qquad \textit{for all } F_1, F_2 \in \mathcal{F}(M,\eta).$$

5.2. Alpha-stable laws. Theorem 5.1 shows that a single set of MIPT thresholds can work not only for Gaussian data but also for a wide family of distributions—provided that we avoid the use of coefficients at the finest scales. To illustrate the theorem, we consider symmetric α-stable laws ($S\alpha S$) [27]. Alpha-stable laws are good models for many applications because of their high variability [24, 27].

Each symmetric α-stable law $S\alpha S$ is specified by its characteristic function $\exp(-\sigma^\alpha|\theta|^\alpha)$, with two parameters, (α,σ), $\alpha \in (0,2]$, $\sigma > 0$. The case $\alpha = 2$ is the Gaussian distribution with standard deviation $\sqrt{2}\sigma$ and density function $1/(\sqrt{2\pi}\sqrt{2}\sigma)\exp(-t^2/(4\sigma^2)$. The case $\alpha = 1$ is the Cauchy distribution with density $\sigma/(\pi(\sigma^2+t^2))$.

For our purposes, we consider $S\alpha S$ densities with σ calibrated so that the density at zero has the same value $1/\sqrt{2\pi}$ as the standard Gaussian. We denote the density and distribution of a $S\alpha S$ standardized in this way by f_α and F_α, respectively. Notice that

$$f_\alpha(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-i\omega t}e^{-\sigma^\alpha|\omega|^\alpha}d\omega = \frac{1}{\pi}\int_0^\infty e^{-\sigma^\alpha\omega^\alpha}\cos(\omega t)d\omega \tag{5.3}$$

and therefore $f_\alpha(0) = \frac{1}{\sigma\pi}I(\alpha)$, where $I(\alpha) = \int_0^\infty e^{-\omega^\alpha}d\omega$. So f_α is properly calibrated by choosing

$$\sigma = \sigma_\alpha = \sqrt{2/\pi}\cdot I(\alpha). \tag{5.4}$$

It is clear from (5.3) that $f_\alpha(t)$ is smooth; the appendix proves the following lemma.

LEMMA 5.2. *Let $0 < \alpha_0 < 2$. $\{F_\alpha : \alpha \in [\alpha_0, 2]\}$ is a subset of a $\mathcal{F}(M,\eta)$ for appropriate M and η.*

Combining this with Theorem 5.1 gives the following corollary.

COROLLARY 5.3. *For any $\epsilon > 0$, $\theta \in (0,1)$ and $\alpha_0 \in (0,2)$, there exists $J^* = J^*(\epsilon,\theta,\alpha_0)$ such that if $J \geq J^*$, then*

$$\max_{j\leq\lfloor\theta J\rfloor}|t_j(F_{\alpha_1}) - t_j(F_{\alpha_2})| \leq \epsilon \qquad \textit{for all}\ \ \alpha_0 \leq \alpha_1,\alpha_2 \leq 2.$$

To illustrate Corollary 5.3, we compare $t_j(F_2)$ with $t_j(F_1)$ in Table 5.1.

While the Gaussian and Cauchy are widely different distributions, their MIPT thresholds are very close at coarse scales.

5.3. Denoising in Gaussian noise. In order to test the above ideas, we first report on the behavior of MIPT with Gaussian noise. Figure 5.1 shows two objects—Heaviside and Doppler—contaminated by Gaussian noise.

The figure also shows the results for (a) Doppler signal, thresholding in MIPT domain; (b) Doppler signal, thresholding in AIPT domain; (c) Heaviside signal, thresholding in MIPT domain; (d) Heaviside signal, thresholding in AIPT domain.

The thresholds t_j in both cases were set by (5.2). The performance of MIPT is comparable to the performance of AIPT, as we expect.

5.4. Denoising in heavy-tailed noise. Next we report on the behavior of MIPT with Cauchy noise. Figure 5.2 shows two objects—Heaviside and Doppler—contaminated by Cauchy noise.

The figure also shows the results for (a) Doppler signal, thresholding in MIPT domain; (b) Doppler signal, thresholding in AIPT domain; (c) Heaviside signal, thresholding in MIPT domain; (d) Heaviside signal, thresholding in AIPT domain.

The thresholds t_j for MIPT thresholding were again set by (5.2). As a control experiment, the same set of thresholds were used for AIPT thresholding. The performance of MIPT is much better than the performance of AIPT, as we expect.

Table 5.1
MIPT thresholds for Gaussian and Cauchy white noise.

N	j	n_j	$t_j(N(0.1))$	$t_j(\text{Cauchy})$	N	j	n_j	$t_j(N(0,1))$	$t_j(\text{Cauchy})$
3^{10}					3^{11}				
	3	2187	6.6286	6.6764		3	6561	6.9052	6.9231
	4	729	6.6306	6.7756		4	2187	6.9059	6.9600
	5	243	6.6366	7.0866		5	729	6.9082	7.0724
	6	81	6.6543	8.1534		6	243	6.9149	7.4261
	7	27	6.7058	12.9767		7	81	6.9350	8.6533
	8	9	6.8393	67.3342		8	27	6.9928	14.4074
	9	3	7.0693	19659.0616		9	9	7.1406	88.0447
	10	1	7.2704	1.4168×10^{12}		10	3	7.3833	43575.6704
						11	1	7.5864	1.5470×10^{13}
3^{12}					3^{13}				
	3	19683	7.1696	7.1763		3	59049	7.4232	7.4257
	4	6561	7.1699	7.1900		4	19683	7.4233	7.4308
	5	2187	7.1707	7.2312		5	6561	7.4237	7.4460
	6	729	7.1732	7.3573		6	2187	7.4246	7.4918
	7	243	7.1808	7.7555		7	729	7.4274	7.6320
	8	81	7.2032	9.1531		8	243	7.4358	8.0765
	9	27	7.2676	15.9479		9	81	7.4606	9.6545
	10	9	7.4294	114.8298		10	27	7.5318	17.6099
	11	3	7.6836	96054.9354		11	9	7.7074	149.4526
	12	1	7.8869	1.6131×10^{14}		12	3	7.9718	210754.2497
						13	1	8.2100	1.5790×10^{15}

6. Appendix: Proofs.

6.1. Preliminaries. Our proofs of P3–P5 rely on two basic facts about medians and median-interpolating refinement.

Lemma 6.1. *Let I be a closed interval and $||f-g||_{L^\infty(I)} \leq \epsilon$, then $|\operatorname{med}(f|I) - \operatorname{med}(g|I)| \leq \epsilon$.*

Proof. $\operatorname{med}(\cdot|I)$ is a monotone functional: $f \leq g \Rightarrow \operatorname{med}(f|I) \leq \operatorname{med}(g|I)$. If $||f-g||_{L^\infty(I)} \leq \epsilon$, then $f \leq g + \epsilon$ and

$$\operatorname{med}(f|I) \leq \operatorname{med}(g+\epsilon\,|I) = \operatorname{med}(g|I) + \epsilon.$$

By symmetry, we also get $\operatorname{med}(g|I) \leq \operatorname{med}(f|I) + \epsilon$. □

Lemma 6.2. *The operators $\Pi_{(2)}, Q_{(2)} : \mathbb{R}^3 \to \mathbb{R}^3$ are Lipschitz operators, i.e., if $\mathbf{m} = (m_1, m_2, m_3)^T$ and $\mathbf{m}' = (m_1', m_2', m_3')^T$, then $||\Pi_{(2)}(\mathbf{m}) - \Pi_{(2)}(\mathbf{m}')||_\infty \leq C \cdot ||\mathbf{m} - \mathbf{m}'||_\infty$ and $||Q_{(2)}(\mathbf{m}) - Q_{(2)}(\mathbf{m}')||_\infty \leq C' \cdot ||\mathbf{m} - \mathbf{m}'||_\infty$.*

Proof. We focus on the proof for $\Pi_{(2)}$; the proof for $Q_{(2)}$ is similar. The closed-form expressions (2.11) and (2.14–2.17) can be used to show that $a(\mathbf{m}), b(\mathbf{m}), c(\mathbf{m})$ are globally continuous, and even that they are analytic within each of the "branches" $\mathcal{N}_1 = \{\mathbf{m} : (m_3 - m_2)/(m_2 - m_1) \in [7/3, 5]\}$, $\mathcal{N}_2 = \{\mathbf{m} : (m_3 - m_2)/(m_2 - m_1) \in [1/5, 3/7]\}$, $\mathcal{N}_3 = \{\mathbf{m} : (m_3 - m_2)/(m_2 - m_1) \in [-3, -1/3]\}$, and $\mathcal{L} = \mathbb{R}^3 - \cup_{i=1}^3 \mathcal{N}_i$. In particular, $a(\mathbf{m}), b(\mathbf{m}), c(\mathbf{m})$ have bounded partial derivatives in each of $\mathcal{N}_i$ and constant partial derivatives in $\mathcal{L}$. Hence, there is a constant $C > 0$ such that if both $\mathbf{m}, \mathbf{m}'$ belong to one of $\mathcal{N}_i$ and $\mathcal{L}$, then

$$|| \left(a(\mathbf{m}), b(\mathbf{m}), c(\mathbf{m})\right)^T - \left(a(\mathbf{m}'), b(\mathbf{m}'), c(\mathbf{m}')\right)^T ||_\infty \leq C \cdot ||\mathbf{m} - \mathbf{m}'||_\infty. \tag{6.1}$$

It remains to show that (6.1) holds also without the restriction that $\mathbf{m}, \mathbf{m}'$ both have to belong to one of $\mathcal{N}_i$ and $\mathcal{L}$. Notice that each $\mathcal{N}_i$ is a *convex* set in $\mathbb{R}^3$ because

$$d_1 \leq \frac{m_3 - m_2}{m_2 - m_1} \leq d_2 \text{ and } d_1 \leq \frac{m_3' - m_2'}{m_2' - m_1'} \leq d_2$$

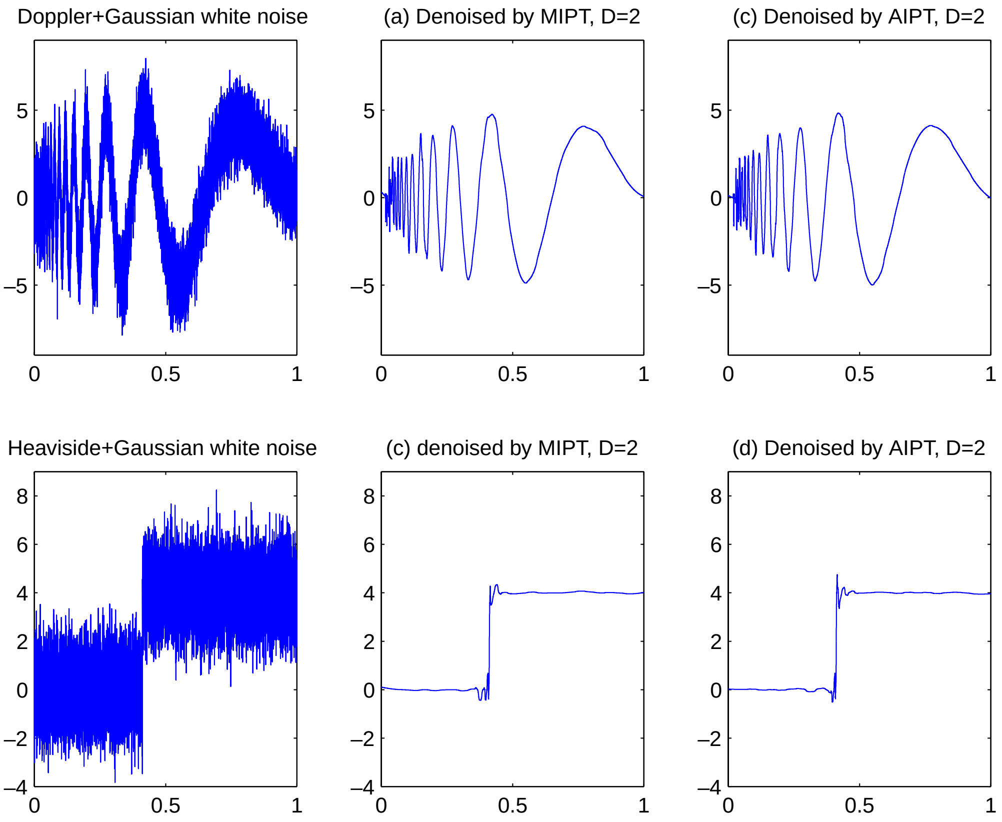


FIG. 5.1. *Denoising of Gaussian data with AIPT and MIPT thresholding.*

implies

$$d_1 \le \frac{(m_3 + t(m_3 - m_3')) - (m_2 + t(m_2 - m_2'))}{(m_2 + t(m_2 - m_2')) - (m_1 + t(m_1 - m_1'))} \le d_2 \quad \text{for all } t \in [0,1].$$

Also for each $d \in \mathbb{R}$, there is a unique $t \in \mathbb{R}$ such that $((m_3 + t(m_3 - m_3')) - (m_2 + t(m_2 - m_2')))/((m_2 + t(m_2 - m_2')) - (m_1 + t(m_1 - m_1'))) = d$. Since $\mathbb{R}$ is the disjoint union of the seven intervals $(-\infty, -3]$, $[-3, -1/3]$, $[-1/3, 1/5]$, $[1/5, 3/7]$, $[3/7, 7/3]$, $[7/3, 5]$, $[5, +\infty)$, we conclude that the line segment (in $\mathbb{R}^3$) joining $\mathbf{m}$ and $\mathbf{m}'$ is the disjoint union of at most seven subsegments each lying completely in one of the sets $\mathcal{N}_i$ and $\mathcal{L}$. Hence by replacing C by $7 \cdot C$ in (6.1) one makes the bound valid for all $\mathbf{m}$ and $\mathbf{m}'$. □

Comment. If $\mathbf{m}$, $\mathbf{m}'$ are associated with triadic intervals $I_{j,k+e}$, $e = -1, 0, 1$ and $\pi_{j,k}$ and $\pi'_{j,k}$ are the corresponding median-interpolants, then Lemma 6.2 also implies that $||\pi_{j,k} - \pi'_{j,k}||_{L^\infty(I_{j,k})} \le C||\mathbf{m} - \mathbf{m}'||_\infty$, where C is an absolute constant independent of $\mathbf{m}$, $\mathbf{m}'$, j, and k.

6.2. Proof of P3. We first recall a standard result in approximation theory.

LEMMA 6.3. *Let $f \in \dot{C}^{r+\alpha}$, $r = 0, 1, 2, \ldots$ and $0 \le \alpha \le 1$. Then there exists a constant C, proportional to the Hölder constant of f, so that for any small enough*

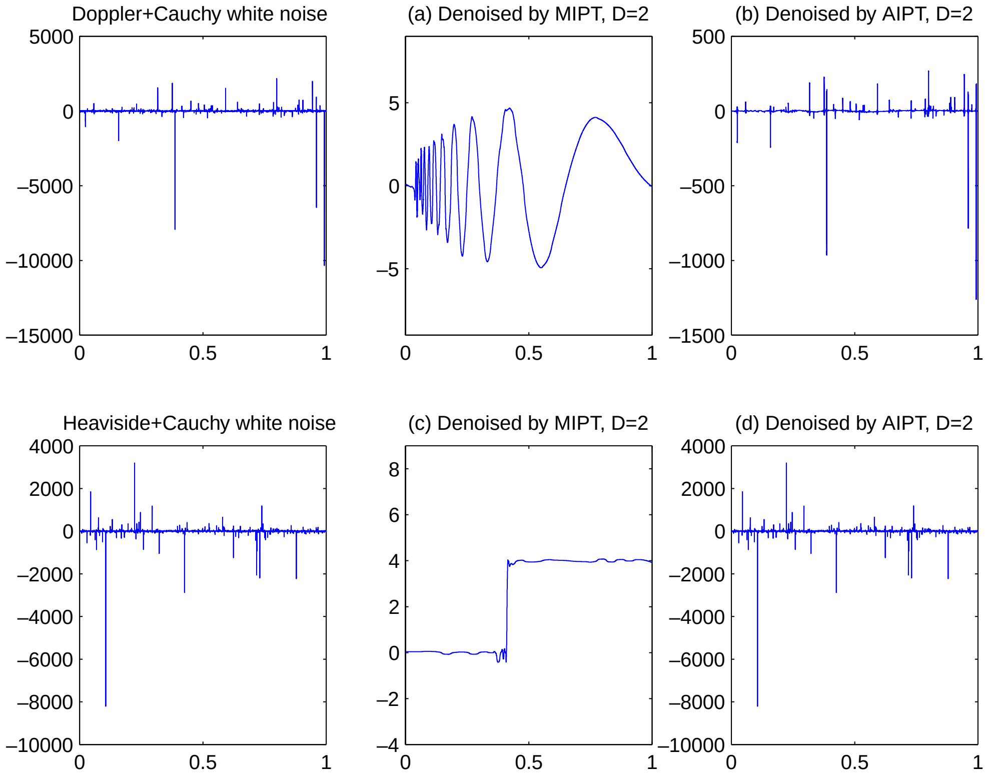


Fig. 5.2. *Denoising of Cauchy data with AIPT and MIPT thresholding.*

interval I, there is a polynomial π_I of degree r with

$$\|f - \pi_I\|_{L^\infty(I)} \le C|I|^{r+\alpha}.$$

Proof of P3. Let $f \in \dot{C}^{r+\alpha}$ ($r = 0, 1$, or 2, $0 \le \alpha \le 1$) and $I_{j,k}$ be an arbitrary triadic interval (with j large enough.) By Lemma 6.3, there exists a degree r polynomial, $\tilde{\pi}_{j,k}$, such that

$$||f - \tilde{\pi}_{j,k}||_{L^\infty(I_{j,k-1}\cup I_{j,k}\cup I_{j,k+1})} \le CC_1 3^{-(r+\alpha)j}. \tag{6.2}$$

Put for short $\epsilon = CC_1 3^{-(r+\alpha)j}$.

Recall the notation $m_{j,k} = \text{med}(f|I_{j,k})$ and let $\pi_{j,k}$ be the degree $D = 2$ polynomial that interpolates the block medians $m_{j,k'}$, $k' = k-1, \dots, k+1$. We want to show that $\pi_{j,k}$ is close to $\tilde{\pi}_{j,k}$. Denote $\tilde{m}_{j,k'} = \text{med}(\tilde{\pi}_{j,k}|I_{j,k'})$, by (6.2) and Lemma 6.1,

$$|m_{j,k'} - \tilde{m}_{j,k'}| \le \epsilon \text{ for } k' = k-1, k, k+1. \tag{6.3}$$

By Lemma 6.2 and (6.3),

$$\left\|\Pi_{(2)}(m_{j,k-1}, m_{j,k}, m_{j,k+1}) - \Pi_{(2)}(\tilde{m}_{j,k-1}, \tilde{m}_{j,k}, \tilde{m}_{j,k+1})\right\|_{L^\infty(I_{j,k})} \le c\epsilon.$$

But $\Pi_{(2)}(m_{j,k-1}, m_{j,k}, m_{j,k+1}) = \pi_{j,k}$ and $\Pi_{(2)}(\tilde{m}_{j,k-1}, \tilde{m}_{j,k}, \tilde{m}_{j,k+1}) = \tilde{\pi}_{j,k}$, hence

$$||f - \pi_{j,k}||_{L^\infty(I_{j,k})} \le ||f - \tilde{\pi}_{j,k}||_{L^\infty(I_{j,k})} + ||\pi_{j,k} - \tilde{\pi}_{j,k}||_{L^\infty(I_{j,k})} \le c''\epsilon. \tag{6.4}$$

Finally, use Lemma 6.1 and (6.4) to conclude that for $e = 0, 1, 2$,

$$|d_{j+1,3k+e}| = |\operatorname{med}(f|I_{j+1,3k+e}) - \operatorname{med}(\pi_{j,k}|I_{j+1,3k+e})| \leq c'''\epsilon$$

or to write it in a cleaner form, $|d_{j,k}| \leq c^{(iv)} \cdot C3^{-(r+\alpha)j}$, where $c^{(iv)} = c'''3^{r+\alpha}$. □

6.3. Proof of P4 and P5. Since

$$\alpha_{j,3k+\epsilon} = m_{j,3k+\epsilon} - (Q_{(2)}(m_{j-1,k-1}, m_{j-1,k}, m_{j-1,k+1}))_\epsilon$$

and $Q_{(2)}$ is Lipschitz (Lemma 6.2), there is a constant $c > 0$, independent of j and k, such that $|\alpha_{j,3k+\epsilon}| \leq c \cdot \max(|m_{j,3k+\epsilon}|, |m_{j-1,k-1}|, |m_{j-1,k}|, |m_{j-1,k+1}|)$. Boole's inequality gives for random variables W_i that

$$P(\max(W_1, \ldots, W_4) > \xi) \leq \sum_{i=1}^{4} P(W_i > \xi)$$

and so we can write

$$P(\sqrt{3^{J-j}}|\alpha_{j,k}| \geq \xi) \leq 4 \cdot P(\sqrt{3^{J-j}}|m_{j,k}| \geq \xi/c). \tag{6.5}$$

Thus, P4 and P5 boil down to the calculation of $P(\sqrt{n}|\operatorname{med}(X_1, \ldots, X_n)| \geq \xi)$ for $X_i \sim_{i.i.d.}$ Gaussian and Cauchy.

We first develop an inequality which derives from standard results in order statistics [6] and in Cramèr–Chernoff bounds on large deviations [7].

LEMMA 6.4. *Let $X_1, \ldots, X_n$ be i.i.d. with cumulative distribution function (c.d.f.) $F(\cdot)$. We have the following estimate:*

$$\begin{aligned} &P\{|(\operatorname{med}(X_1, \ldots, X_n)| \geq x\} \\ &\qquad \leq \min\left\{1, \left[2\sqrt{F(x)(1-F(x))}\right]^n + \left[2\sqrt{F(-x)(1-F(-x))}\right]^n\right\}. \end{aligned}$$

Proof. It suffices to show that $P(\operatorname{med}(X_1, \ldots, X_n) \geq x) \leq \left[2\sqrt{F(x)(1-F(x))}\right]^n$ for any $x \geq 0$. Let $I_i = 1_{(X_i \geq x)}$,

$$P\left(\operatorname{med}(X_1, \ldots, X_n) \geq x\right) \leq P\left(\sum_{i=1}^{n} I_i \geq \frac{n}{2}\right).$$

Since $I_i \sim_{\text{i.i.d.}} \text{Binomial}(1, 1-F(x))$, $S_n := \sum_{i=1}^n I_i \sim \text{Binomial}(n, 1-F(x))$. By $1_{(S_n \geq \frac{n}{2})} \leq e^{\lambda S_n} e^{-\lambda\frac{n}{2}}$ for all $\lambda > 0$, we have

$$\begin{aligned} P(S_n \geq \frac{n}{2}) &\leq \min_{\lambda>0} \mathrm{E}(e^{\lambda S_n} e^{-\lambda\frac{n}{2}}) = \min_{\lambda>0} e^{-\lambda\frac{n}{2}} \mathrm{E}(e^{\lambda\sum_1^n I_i}) \\ &= \min_{\lambda>0} e^{-\lambda\frac{n}{2}} \left(F(x) + (1-F(x))e^{\lambda}\right)^n = \left[\min_{\lambda>0} F(x)e^{-\frac{\lambda}{2}} + (1-F(x))e^{\frac{\lambda}{2}}\right]^n \\ &= \left(F(x)e^{-\frac{\lambda}{2}} + (1-F(x))e^{\frac{\lambda}{2}}|_{\lambda=\ln\left(\frac{F(x)}{1-F(x)}\right)}\right)^n = \left(2\sqrt{F(x)(1-F(x))}\right)^n. \quad \square \end{aligned}$$

COROLLARY 6.5. *Let $X_1, \ldots, X_n$ be i.i.d. with c.d.f. $F(x)$ having symmetric density f. Given $\alpha \in (0, 1/2)$, define*

$$t_{\alpha,n} := F^{-1}\left(\frac{1}{2} + \frac{1}{2}\sqrt{1 - \left(\frac{\alpha}{2}\right)^{\frac{2}{n}}}\right),$$

then

$$P\left(|\mathrm{med}(X_1,\dots,X_n)| \ge t_{\alpha,n}\right) \le \alpha.$$

We now apply Lemma 6.4 to the Gaussian and Cauchy distributions. Since they are both symmetric distributions, we have

$$\begin{aligned} P\left(\left|\sqrt{n}\ \mathrm{med}(X_1,\dots,X_n)\right| \ge \xi\right) &\le 2\cdot 2^n \sqrt{F\left(\frac{\xi}{\sqrt{n}}\right)\left(1-F\left(\frac{\xi}{\sqrt{n}}\right)\right)}^{\,n} \\ &= 2\cdot\left(\left[4F\left(y\right)\left(1-F\left(y\right)\right)\right]^{y^{-2}/2}\right)^{\xi^2} \\ &= 2\cdot\exp(\theta(y)\xi^2), \end{aligned}$$

where $y \equiv \xi/\sqrt{n}$ and $\theta(y) \equiv y^{-2}/2\cdot\log\left[4F\left(y\right)\left(1-F\left(y\right)\right)\right]$. Gaussian-type probability bounds will follow for a range $0 \le \xi \le X$ in P4 and P5, from an inequality $\sup_{[0,Y]}\theta(y) < 0$ on a corresponding range of values $0 \le y \le Y$, with $Y = X/\sqrt{n}$.

(i) Gaussian distribution: To establish P4, we need inequalities valid for $0 \le \xi < \infty$, i.e., $0 \le y < \infty$. Now

$$\sup_{y\in[2,\infty)} \theta(y) = \sup_{y\in[2,\infty)} y^{-2}/2\cdot\left[\log(1-F_2(y)) + \log(4F_2(y))\right].$$

From Mills' ratio $\int_y^\infty e^{-x^2/2}dx \le \frac{1}{y}e^{-y^2/2}$ holding for all $y > 1$, we have $\log(1-F_2(y)) \le -y^2/2 - \log(y)$. Hence

$$\sup_{y\in[2,\infty)} \theta(y) \le \sup_{y\in[2,\infty)} y^{-2}/2\cdot\left[-y^2/2 - \log(y) + \log(4)\right] = -(2-\log(2))/2 < 0.$$

On the other hand, from symmetry of F_2 and unimodality of the density f_2 we get $4F_2(y)(1-F_2(y)) = 4F_2(y)F_2(-y) \le 1 - cy^2$ on $|y| \le 2$, with $c > 0$; so

$$\sup_{y\in[0,2]} \theta(y) = \sup_{y\in[0,2]} y^{-2}/2\cdot\log(1-cy^2) \le -c/2.$$

(ii) Cauchy distribution: $F_1(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\sqrt{\frac{\pi}{2}}x\right)$. To get P5 we aim only for an inequality valid on $y \in [0,Y]$, with $Y = 1$, which gives a Gaussian-type inequality for $\xi \in [0,\sqrt{n}]$.

$$\theta(y) = y^{-2}/2\cdot\log\left[1 - \frac{4}{\pi^2}\arctan^2\left(\sqrt{\frac{\pi}{2}}y\right)\right] \le -y^{-2}/2\cdot\frac{4}{\pi^2}\arctan^2\left(\sqrt{\frac{\pi}{2}}y\right).$$

However, as $\arctan(y) > c\cdot y$ for $y \in [0,1]$, with $c > 0$, this gives $\theta(y) < -\frac{c^2}{\pi}$ for $y \in [0,1]$.

6.4. Proof of Theorem 5.1. Let $p = 1/(2J3^J)^2$, $n_j = 3^{J-j}$. Let j be chosen such that

$$1/2\sqrt{1-p^{1/n_j}} \le \eta. \tag{6.6}$$

Then there exist $0 < \eta_1, \eta_2 \le \eta$ such that

$$\begin{aligned} |t_j(F_1) - t_j(F_2)| &= \sqrt{n_j}\left|F_1^{-1}\left(\frac{1}{2}+\frac{1}{2}\sqrt{1-p^{\frac{1}{n_j}}}\right) - F_2^{-1}\left(\frac{1}{2}+\frac{1}{2}\sqrt{1-p^{\frac{1}{n_j}}}\right)\right| \\ &= \sqrt{n_j}\left|(F_1^{-1})''(1/2+\eta_1) - (F_2^{-1})''(1/2+\eta_2)\right|\left(\frac{1}{2}\sqrt{1-p^{\frac{1}{n_j}}}\right)^2 \\ &\le M/2\sqrt{n_j}\left(1-p^{\frac{1}{n_j}}\right), \end{aligned}$$

i.e., $|t_j(F_1) - t_j(F_2)| \leq \epsilon$ if

$$\sqrt{n_j}\left(1 - p^{\frac{1}{n_j}}\right) \leq \frac{2\epsilon}{M}. \tag{6.7}$$

Since $(1-x) \leq \log(1/x)$ for all $x \in (0,1]$, (6.6) holds for large enough J because

$$\begin{aligned}\frac{1}{2}\sqrt{1 - p^{1/n_j}} \leq \frac{1}{2}\sqrt{\log p^{-1/n_j}} &= \frac{1}{2}\sqrt{\frac{1}{n_j}\log(4 \cdot J^2 \cdot 3^{2J})} \\ &\leq \frac{1}{2}\sqrt{\frac{1}{3^{(1-\theta)J}}(\log(4) + 2\log(J) + 2J\log(3))} \to 0 \text{ as } J \to 0.\end{aligned}$$

Similarly, (6.7) holds for large enough J because

$$\begin{aligned}\sqrt{n_j}\left(1 - p^{\frac{1}{n_j}}\right) &\leq (1/\sqrt{n_j})\log p^{-1} \\ &\leq \frac{1}{\sqrt{3^{(1-\theta)J}}}(\log(4) + 2\log(J) + 2J\log(3)) \to 0 \text{ as } J \to 0. \qquad \square\end{aligned}$$

6.5. Proof of Lemma 5.2. It suffices to find M, η such that

$$\sup_{\alpha \in [\alpha_0, 2]} \sup_{p \in [\frac{1}{2}-\eta, \frac{1}{2}+\eta]} (F_\alpha^{-1})''(p) \leq M.$$

Since

$$(F_\alpha^{-1})''(p) = -\frac{f_\alpha'(F_\alpha^{-1}(p))}{[f_\alpha(F_\alpha^{-1}(p))]^3},$$

we work with F_α^{-1}, f_α, and f_α' separately.

1. Since $|F_\alpha^{-1}(p)|$ is monotone increasing in p for fixed α, and is monotone increasing in α for fixed p, we have, for any $0 < \eta < 1/2$,

$$\sup_{0<\alpha\leq 2} \sup_{1/2-\eta \leq p \leq 1/2+\eta} |F_\alpha^{-1}(p)| = F_2^{-1}(1/2 + \eta).$$

2. Now $\sup_{|t|\leq\epsilon_1} |f_\alpha(t) - f_\alpha(0)| \leq |t| \cdot \{\sup_{|t|\leq\epsilon_1} |f_\alpha'(t)|\}$. Also

$$\begin{aligned}\sup_{|t|\leq\epsilon_1} |f_\alpha'(t)| &= \sup_{|t|\leq\epsilon_1}\left|\frac{1}{\pi}\int_0^\infty e^{-\sigma_\alpha^\alpha\omega^\alpha}(-\omega)\sin(\omega t)d\omega\right| \\ &\leq \frac{1}{\pi}\int_0^\infty e^{-\sigma_\alpha^\alpha\omega^\alpha}\omega d\omega = \frac{1}{\pi}\frac{1}{\sigma_\alpha^2}\int_0^\infty e^{-\omega^\alpha}\omega d\omega \leq C_1(\alpha_0),\end{aligned}$$

where $C_1(\alpha) = \frac{1}{2}(\int_0^\infty e^{-\omega^\alpha}\omega d\omega)/(\int_0^\infty e^{-\omega^\alpha}d\omega)^2$ is defined on $(0,2]$ and is positive and monotone decreasing.

3. Similarly $\sup_{|t|\leq\epsilon_1} |f_\alpha'(t) - f_\alpha'(0)| \leq |t| \cdot \{\sup_{|t|\leq\epsilon_1} |f_\alpha''(t)|\}$. Moreover

$$\begin{aligned}\sup_{|t|\leq\epsilon_1} |f_\alpha''(t)| &= \sup_{|t|\leq\epsilon_1}\left|\frac{1}{\pi}\int_0^\infty e^{-\sigma_\alpha^\alpha\omega^\alpha}(\omega^2)\cos(\omega t)d\omega\right| \\ &\leq \frac{1}{\pi}\int_0^\infty e^{-\sigma_\alpha^\alpha\omega^\alpha}\omega^2 d\omega = \frac{1}{\pi}\frac{1}{\sigma_\alpha^3}\int_0^\infty e^{-\omega^\alpha}\omega^2 d\omega \leq C_2(\alpha_0),\end{aligned}$$

where $C_2(\alpha) = \frac{1}{2}\sqrt{\frac{\pi}{2}}(\int_0^\infty e^{-\omega^\alpha}\omega^2 d\omega)/(\int_0^\infty e^{-\omega^\alpha} d\omega)^3$ is defined on $(0, 2]$ and is positive and monotone decreasing. Therefore

$$\begin{aligned}\sup_{\alpha\in[\alpha_0,2]} \sup_{p\in[\frac{1}{2}-\eta,\frac{1}{2}+\eta]} |(F_\alpha^{-1})''(p)| &= \sup_{\alpha\in[\alpha_0,2]} \sup_{p\in[\frac{1}{2}-\eta,\frac{1}{2}+\eta]} \frac{|f'_\alpha(F_\alpha^{-1}(p))|}{|f_\alpha(F_\alpha^{-1}(p))|^3} \\ &\le \frac{|F_2^{-1}(\frac{1}{2}+\eta)|C_2(\alpha_0)}{\left|\frac{1}{\sqrt{2\pi}} - F_2^{-1}(\frac{1}{2}+\eta)C_1(\alpha_0)\right|^3}. \qquad (6.8)\end{aligned}$$

If we choose $\eta = \eta(\alpha_0) > 0$ small enough such that $F_2^{-1}(\frac{1}{2}+\eta)C_1(\alpha_0) < \frac{1}{\sqrt{2\pi}}$, then with the choice of $M = M(\alpha_0)$ defined by (6.8) we get $\{F_\alpha : \alpha_0 \le \alpha \le 2\} \subset \mathcal{F}(M(\alpha_0), \eta(\alpha_0))$. □

Reproducible research. In this paper, all computational results are reproducible, meaning that the code which generated the figures is available over the Internet, following the discipline indicated in [3]. Interested readers are directed to http://www-stat.stanford.edu/˜wavelab/.

Acknowledgments. The authors would like to thank Andrew Bruce, Amir Dembo, Gary Hewer, and Charles Micchelli for helpful discussions and references.

REFERENCES

[1] B. Borden and M. Mumford, *A statistical glint/radar cross section target model*, IEEE Trans. Aerospace Electron. Systems, 19 (1983), pp. 781–785.

[2] A. G. Bruce, D. L. Donoho, H.-Y. Gao, and R. D. Martin, *Denoising and robust nonlinear wavelet analysis*, in SPIE Proceedings, Wavelet Appl. 2242, Orlando, FL, 1994.

[3] J. Buckheit and D. L. Donoho, *Wavelab and reproducible research*, in Wavelets in Statistics, A. Antoniadis and G. Oppenheim, eds., Springer-Verlag, New York, 1994, pp. 55–82.

[4] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math., SIAM, Philadelphia, 1992.

[5] I. Daubechies and J. Lagarias, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[6] H. A. David, *Order Statistics*, Wiley, New York, London, Sydney, 1970.

[7] A. Dembo and O. Zeitouni, *Large Deviations and Applications*, Bartlett and Jones, Boston, 1993.

[8] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.

[9] D. L. Donoho, *Interpolating Wavelet Transforms*, tech. report, Department of Statistics, Stanford University, Stanford, CA, 1992. Available at http://www-stat.stanford.edu/˜donoho/Reports/1992/interpol.ps.Z

[10] D. L. Donoho, *Smooth wavelet decompositions with blocky coefficient kernels*, in Recent Advances in Wavelet Analysis, L. Schumaker and G. Webb, eds., Academic Press, Boston, 1993, pp. 259–308.

[11] D. L. Donoho, *Minimum entropy segmentation*, in Wavelets Theory, Algorithms and Applications, C. Chui, L. Montefusco, and L. Puccio, eds., Academic Press, San Diego, 1994, pp. 233–269.

[12] D. L. Donoho, *De-noising by soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–27.

[13] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness by wavelet shrinkage*, J. Amer. Statist. Assoc., 90 (1995), pp. 1200–1224.

[14] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 301–369.

[15] D. L. Donoho and T. P.-Y. Yu, *Nonlinear "Wavelet Transforms" Based on Median-Interpolation*, tech. report, Department of Statistics, Stanford University, Stanford, CA, 1997. Available at http://www-stat.stanford.edu/˜donoho/Reports/1997/median.ps.Z.

[16] N. Dyn, J. Gregory, and D. Levin, *Analysis of uniform binary subdivision schemes for curve design*, Constr. Approx., 7 (1991), pp. 127–147.

[17] H.-Y. Gao., *Wavelet Estimation of Spectral Densities in Time Series Analysis*, Ph.D. Thesis, Department of Statistics, University of California, Berkeley, CA, 1993.

[18] T. N. T. Goodman and T. P.-Y. Yu, *Interpolation of medians*, Adv. Comput. Math., 11 (1999), pp. 1–10.

[19] D. Greer, I. Fung, and J. Shapiro, *Maximum-likelihood multiresolution laser radar range imaging*, IEEE Trans. Image Process., 6 (1997), pp. 36–47.

[20] P. Hall and P. Patil, *On the choice of smoothing parameter, threshold and truncation in nonparametric regression by wavelet methods*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 361–377.

[21] F. R. Hampel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, 1986.

[22] P. Huber, *Robust Statistics*, Wiley, New York, 1981.

[23] H. Longbotham, *A class of robust nonlinear filters for signal decomposition and filtering utilizing the Haar basis*, in ICASSP-92, Vol. 4, IEEE Signal Processing Society, Piscataway, NJ, 1992.

[24] C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*, Wiley, New York, 1995.

[25] O. Rioul, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.

[26] D. F. F. R. L. de Queiroz and R. Schafer, *Nonexpansive pyramid for image coding using a non-linear filter bank*, IEEE Trans. Image Process., 7 (1998), pp. 246–252.

[27] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Process: Stochastic Models with Infinite Variance*, Chapman & Hall, New York, 1994.

[28] J. L. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analysis: The Multiscale Approach*, Cambridge University Press, Cambridge, UK, 1998.

[29] J. Starck, F. Murtagh, B. Pirenne, and M. Albrecht, *Astronomical image compression based on noise suppression*, Proc. Astro. Soc. Pac., 108 (1996), pp. 446–455.

[30] B. Stuck and B. Kleiner, *A statistical analysis of telephone noise*, Bell System Tech. J., 53 (1974), pp. 1263–1320.

[31] T. P.-Y. Yu, *New Developments in Interpolating Wavelet Transforms*, Ph.D. thesis, Program of Scientific Computing and Computational Mathematics, Stanford University, Stanford, CA, 1997.

SIAM J. MATH. ANAL.
Vol. 31, No. 5, pp. 1062–1099

ORTHONORMAL RIDGELETS AND LINEAR SINGULARITIES[*]

DAVID L. DONOHO[†]

Abstract. We construct a new orthonormal basis for $L^2(\mathbb{R}^2)$, whose elements are angularly integrated ridge functions—*orthonormal ridgelets*. The basis elements are smooth and of rapid decay in the spatial domain, and in the frequency domain are localized near angular wedges which, at radius $r = 2^j$, have radial extent $\Delta r \approx 2^j$ and angular extent $\Delta\theta \approx 2\pi/2^j$.

Orthonormal ridgelet expansions expose an interesting phenomenon in nonlinear approximation: they give very efficient approximations to objects such as $1_{\{x_1 \cos\theta + x_2 \sin\theta > a\}}\ e^{-x_1^2 - x_2^2}$ which are smooth away from a discontinuity along a line. The orthonormal ridgelet coefficients of such objects are *sparse*: they belong to every ℓ^p, $p > 0$. This implies that simple thresholding in the ridgelet orthobasis is, in a certain sense, a near-ideal nonlinear approximation scheme for such objects.

Orthonormal ridgelets may be viewed as L^2 substitutes for approximation by sums of ridge functions, and so can perform many of the same tasks as the ridgelet systems constructed by Candès [Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, CA, 1998; *Appl. Comput. Harmon. Anal.*, 6 (1999), pp. 197–218]. Orthonormal ridgelets make available the machinery of orthogonal decompositions, which is not available for ridge functions as they are not in $L^2(\mathbb{R}^2)$.

The ridgelet orthobasis is constructed as the isometric image of a special wavelet basis for Radon space; as a consequence, ridgelet analysis is equivalent to a special wavelet analysis in the Radon domain. This means that questions of ridgelet analysis of linear singularities can be answered by wavelet analysis of point singularities. At the heart of our nonlinear approximation result is the study of a certain tempered distribution on $\mathbb{R}^2$ defined formally by $S(u, v) = |v|^{-1/2}\sigma(u/|v|)$ with σ a certain smooth bounded function; this is singular at $(u, v) = (0, 0)$ and C^∞ elsewhere. The key point is that the analysis of this point singularity by tensor Meyer wavelets yields sparse coefficients at high frequencies; this is reflected in the sparsity of the ridgelet coefficients and the good nonlinear approximation properties of the ridgelet basis.

Key words. wavelets, singularities, edges, ridge function, ridgelet, radon transform, nonlinear approximation, thresholding of wavelet coefficients

AMS subject classifications. 41A63, 41A25, 33E99

PII. S0036141098344403

1. Introduction.

1.1. Sparse representation of singularities.

One of the most striking features of wavelet analysis is its ability to efficiently represent functions which are smooth away from *point singularities.* To see what we mean, consider the function $f_\alpha(x) = |x|^{-\alpha}w(x)$ of $x \in \mathbb{R}^2$, where $w(x)$ is a smooth window of compact support and $\alpha < 1/2$. Now f is smooth away from 0 and has a square-integrable singularity at the point $x = 0$. The coefficients of f in the Meyer orthonormal wavelet basis are *sparse*: arranging them in decreasing order of magnitude gives a sequence decaying more rapidly than any negative power of the index. In this regard, the wavelet coefficients of a point singularity behave similarly to the wavelet coefficients of a smooth function (such as $w(x)$); the sparsity of a wavelet analysis is in a sense insensitive to the presence of point singularities.

Sparsity of the wavelet coefficients has implications for the quality of partial wavelet reconstructions. If we approximate a function using just the m-best terms in

[*] Received by the editors September 10, 1998; accepted for publication July 28, 1999; published electronically April 20, 2000. This research was supported by National Science Foundation grant DMS 95–05151 and by AFOSR MURI 95–P49620–96–1–0028.

http://www.siam.org/journals/sima/31-5/34440.html

[†] Department of Statistics, Stanford University, Stanford, CA 94305-4065 (donoho@stat.stanford.edu).

the wavelet expansion, and if the coefficients are sparse in the sense just given, then the L^2 error of best-m-term approximation decays rapidly with m—faster than any negative power of m. Hence, the fact that wavelet analysis of a point singularity yields sparse coefficients means that smooth functions with point singularities can be very efficiently approximated by partial wavelet reconstructions. This fact has significant implications in data compression and in statistical estimation. (Extensive references on these implications are given in [6], [7].

Point singularities are just one possible type of singularity. Consider the Gaussian-windowed halfspace

$$g^0(x_1, x_2) = 1_{\{x_2>0\}}\, e^{-x_1^2-x_2^2}, \qquad x \in \mathbb{R}^2. \tag{1.1}$$

This has a singularity along the line $x_2 = 0$. One can also consider the more general family

$$g(x_1, x_2; \theta^0, x^0) = 1_{\{x_1\cos(\theta^0)+x_2\sin(\theta^0)>t^0\}} e^{-x_1^2-x_2^2}, \tag{1.2}$$

where $t^0 = x_1^0 \cos(\theta^0) + x_2^0 \sin(\theta^0)$. These functions have a discontinuity along the line $t^0 = x_1 \cos(\theta^0) + x_2 \sin(\theta^0)$ and are smooth elsewhere.

For typical functions of the type (1.1)–(1.2), wavelets do *not* yield sparse coefficients as they did with $f_{0,\alpha}$. For example, in $\mathbb{R}^2$, an object of type g is easily seen to have typically at least order $O(2^j)$ standard wavelet coefficients with amplitude exceeding 2^{-j}. So the mth largest wavelet coefficient of such an object is often of size $\geq c \cdot m^{-1}$ for $c > 0$; this is much poorer decay than what we saw earlier in the case of point singularities, where the decay was faster than any negative power of m. In consequence, m-term wavelet reconstructions do not approximate such objects with the kind of efficiency we saw earlier in the case of point singularities. We can formulate this conclusion more boldly by saying that *wavelets do not efficiently approximate edges in* $\mathbb{R}^2$.

Similar conclusions are possible for Fourier methods. If we consider a function f which is compactly supported in $[0, 2\pi)^2$ and which is smooth away from a linear singularity, and we use the standard bivariate Fourier series to approximate f, we get order $m^{4/3}$ coefficients larger than c/m. So Fourier methods give coefficients which are even less sparse than wavelet coefficients. We can again formulate this conclusion more boldly by saying that *Fourier methods also do not efficiently approximate edges in* $\mathbb{R}^2$.

Observations such as these—and the relative ubiquity of edges in certain applications (such as image processing)—point to the need for better systems of harmonic analysis, ones which efficiently deal with edges, or in another terminology, transforms for which objects like g have sparse coefficients.

1.2. Orthonormal ridgelets. In this article, we introduce a new basis for functions in $L^2(\mathbb{R}^2)$: the *orthonormal ridgelets*, defined as follows. Let $(\psi_{j,k}(t) : j \in \mathbb{Z}, k \in \mathbb{Z})$ be an orthonormal basis of Meyer wavelets for $L^2(\mathbb{R})$ [14], and let $(w^0_{i_0\ell}(\theta),\ \ell = 0, \ldots, 2^{i_0}-1;\ w^1_{i,\ell}(\theta),\ i \geq i_0,\ \ell = 0, \ldots, 2^i-1)$ be an orthonormal basis for $L^2[0, 2\pi)$ made of periodized Lemarié scaling functions $w^0_{i_0\ell}$ at level i_0 and periodized Meyer wavelets $w^1_{i\ell}$ at levels $i \geq i_0$. (We suppose a particular normalization of these functions given in (2.8) below.) Let $\hat{\psi}_{j,k}(\omega)$ denote the Fourier transform of $\psi_{j,k}(t)$, and define ridgelets $\rho_\lambda(x)$, $\lambda = (j, k; i, \ell, \varepsilon)$ as functions of $x \in \mathbb{R}^2$ using the frequency-domain definition

$$\hat{\rho}_\lambda(\xi) = |\xi|^{-\frac{1}{2}} (\hat{\psi}_{j,k}(|\xi|) w^\varepsilon_{i,\ell}(\theta) + \hat{\psi}_{j,k}(-|\xi|) w^\varepsilon_{i,\ell}(\theta + \pi))/2. \tag{1.3}$$

Here the indices run as follows: $j, k \in \mathbb{Z}$, $\ell = 0, \ldots, 2^{i-1} - 1$; $i \geq i_0$, and, if $\varepsilon = 0$, $i = \max(i_0, j)$, while if $\varepsilon = 1$, $i \geq \max(i_0, j)$. Notice the restrictions on the range of i, ℓ. Let Λ denote the set of all such indices λ.

Sections 2 and 3 below establish the following theorem.

Theorem 1.1. *$(\rho_\lambda)_{\lambda \in \Lambda}$ is a complete orthonormal system for $L^2(\mathbb{R}^2)$.*

Define now $\psi^+_{j,k}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\omega|^{\frac{1}{2}} \hat{\psi}_{j,k}(\omega)\ e^{i\omega t} d\omega$; this is a fractionally differentiated Meyer wavelet. Section 4 below shows the following theorem.

Theorem 1.2. *For $x = (x_1, x_2) \in \mathbb{R}^2$,*

$$\rho_\lambda(x) = \frac{1}{4\pi} \int_0^{2\pi} \psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta) d\theta. \tag{1.4}$$

Each $\psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta)$ is a *ridge* function of $x \in \mathbb{R}^2$, i.e., a function of the form $r(x_1 \cos\theta + x_2 \sin\theta)$ [13]. Therefore ρ_λ is obtained by "averaging" ridge functions with ridge angles θ localized near $\theta_{i,\ell} = 2\pi\ell/2^i$; this justifies the "ridgelet" appellation.

1.3. Singularities along lines. Our purpose in this article is to show that orthonormal ridgelets yield efficient representation of objects with linear singularities. We prove in section 5 below the following result for the Gaussian-windowed halfspace $g = g^0$ defined in (1.1).

Theorem 1.3. *The number of ridgelet coefficients of g with amplitude exceeding $1/N$ grows with N more slowly than any fractional power of N.*

In section 6 we consider the general family of smooth functions with linear singularities (1.2). We show that the conclusion of Theorem 1.3 holds for every $g = g(\cdot, \cdot; \theta^0, x^0)$, where $x^0 \in \mathbb{R}^2$ and $\theta^0 \in [0, 2\pi)$. In short, the ridgelet coefficients of such g are *sparse.*

This striking phenomenon may convince the reader that wavelets and other traditional harmonic analysis tools can be substantially improved on as soon as one leaves the setting of pure point singularities.

1.4. Significance for nonlinear approximation and data compression. We now briefly describe the importance of the sparsity of ridgelet expansions for nonlinear approximation and data compression. In effect, the sparsity phenomenon means that a very few select terms will provide good approximation and a good encoding can be had with relatively few bits. Here "relatively few" means as compared with similar encoding schemes based on wavelet or Fourier coefficients.

Consider the following simple nonlinear approximation scheme. Let $\eta_\delta(y) = y 1_{\{|y|>\delta\}}$ denote the *hard thresholding nonlinearity* with threshold $\delta > 0$. With g one of the functions (1.2), define

$$\tilde{g}_\delta = \sum_\Lambda \eta_\delta(\langle g, \rho_\lambda \rangle) \rho_\lambda.$$

This is a finite sum of ridgelets, with $N(\delta) = \sum_\Lambda 1_{\{|\langle g, \rho_\lambda \rangle| > \delta\}}$ terms, built from those ridgelets having large coefficients only. As the threshold $\delta \to 0$, more terms enter the sum, and in general $N(\delta) \uparrow \infty$. Define a sequence of approximants $(\overline{g}_N)$ via $\overline{g}_{N(\delta)} = \tilde{g}_\delta$. These are nonlinear approximants, since the terms which enter in the sum are those which survive thresholding, and this depends on g.

The approximants $(\overline{g}_N)$ converge rapidly to g; from Theorem 1.3, one can see that for each $m > 0$, there is a constant C_m with

$$\|g - \overline{g}_N\|_{L^2(\mathbb{R}^2)} \leq C_m N^{-m}, \qquad N \to \infty.$$

In other words, ridgelets achieve an unlimited rate of approximation. We interpret this by saying that nonlinear ridgelet approximations behave very well for functions which are piecewise smooth, where the boundary between pieces is a line. In comparison, nonlinear wavelet approximations have similar rapid convergence properties for functions which have a punctuated smoothness, i.e., functions which are C^∞ away from isolated point singularities. However, wavelets do not exhibit similarly rapid convergence on objects with discontinuities along linear boundaries, requiring N coefficients to get $O(N^{-1})$ approximation in mean-square.

This fact can be significant for data compression as well. By simply encoding the significant ridgelet coefficients with finite accuracy approximations to the coefficients, and encoding the positions of the coefficients economically, one obtains a finite-precision ridgelet representation of an object; when that object is smooth except for singularities across linear boundaries this representation uses many fewer bits than would be required in wavelet encoding of similar precision. Simple calculations reveal that the number of bits required for ridgelet encoding to accuracy ε in mean square is nearly logarithmic in ε^{-1} as $\varepsilon \to 0$, while the number of bits required for wavelet encoding grows like ε^{-1}. The dramatic difference in growth between $\log(\varepsilon^{-1})$ and ε^{-1} as $\varepsilon \to 0$ quantifies explicitly the extent to which ridgelet methods are better suited towards dealing with edges along linear boundaries.

Ultimately, of course, one wants to go beyond the representation of objects with singularities along lines, to consider the representation of objects with singularities along curves. While ridgelets are essentially focused on dealing with straight lines rather than curves, ridgelets can be adapted to representing objects with curved edges using an appropriate multiscale localization. In effect, multiscale ridgelet expansions divide the image into small dyadic pieces on which the curved edges are nearly straight and uses ridgelet expansions on those pieces. This can be used to obtain expansions of objects with discontinuities along curved edges having significantly more rapid decay of coefficients than the traditional wavelet and Fourier methods. For further discussion on applications see [3].

1.5. On the ridgelet concept. The orthonormal ridgelets are in $L^2(\mathbb{R}^2)$ and so are to be distinguished from the approximation system called ridgelets in the pioneering work of Candès [2], [1]. In Candès' work, the phrase "ridgelet" refers specifically to a ridge function $\psi_{a,b,\theta}(x) = \psi(a(x_1\cos(\theta) + x_2\sin(\theta)) - b)a^{1/2}$, where ψ is oscillatory. As $\psi_{a,b,\theta}$ is constant along "ridges" $x_1\cos(\theta) + x_2\sin(\theta)) = \text{Const}$ and so cannot belong to L^2. This fact creates certain difficulties of construction and interpretation. In order to obtain a method of series representation, Candès constructs "ridgelet frames," where the individual elements of the so-called primal frame have the ridge structure, for appropriate (a_n, b_n, θ_n). This frame construction assumed that the object to be analyzed was supported in a compact set D; frame bounds were established under this condition which proved implicitly the existence of a dual synthesis system in $L^2(D)$. Unfortunately, the construction of dual frames was implicit, depending, for example, on the assumed D, and the properties of the dual frame elements were not available directly, and so it was unclear how to obtain any substantial insight about the structure of the frame expansions. Thus one had a system with analyzing functions of known form, but synthesizing functions of little-known form. Also, owing to the lack of orthogonality, it was unclear how to make good m-term approximations built from such frames.

In the present article we adopt a modified notion of ridgelet, abandoning insistence on the ridge-function form for the elements of the analyzing system, and instead taking the viewpoint that ridgelets should be characterized by certain localization properties they obey in a radial frequency × angular-frequency domain. The formula (1.3) shows that orthonormal ridgelets are localized in the frequency domain into elongated wedges. In polar coordinates these wedges have radial extent 2^j and angular width $2\pi/2^i$ for $i \geq j$. Under this modified notion of ridgelet, this paper shows that it is possible to explicitly construct orthonormal ridgelet bases in $L^2(\mathbb{R}^2)$, in effect, with basis functions available using the simple formula (1.3).

The orthonormal approach has the benefit that analysis functions and synthesis functions are identical and are smooth functions of rapid decay. Moreover, orthonormal ridgelets do not require the object being analyzed to have compact support. The orthonormal approach also has benefits for allowing simple, effective nonlinear approximation. Best-m-term nonlinear approximation in an orthonormal system can be based on simple thresholding ideas. A further benefit is the fact that the orthonormal ridgelet coefficients can be identified with the properties of bivariate wavelet analysis of a fractionally differentiated Radon transform. This identification is used below to show that orthonormal ridgelet coefficients are sparse when analyzing certain smooth objects with linear singularities in $\mathbb{R}^2$.

Throughout this paper, the term ridgelets refers to the system of orthonormal ridgelets introduced here, and not to ridgelet analysis based on ridge functions. We believe that our use of the ridgelets name is sensible because we can show close connections of orthonormal ridgelets to the original ridge function concept. Theorem 1.2 is an instance of this connection; see section 4 below. The companion paper [9] explores carefully the connection between orthonormal ridgelets and ridge functions and gives results showing that orthonormal ridgelets are an effective substitute for ridge function approximation. In effect, that paper shows that a ridge function $\psi^+_{j,k}(x_1\cos(\theta) + x_2\sin(\theta))$ can be approximated with high accuracy $1/N^m$ on a disk of fixed radius using a number of orthonormal ridgelets which grows basically logarithmically in N. In short, "a ridge function is essentially a sum of a few orthonormal ridgelets." This further justifies our use of the ridgelet appellation to label the basis constructed here.

1.6. Wavelet analysis in Radon space. A key structural feature in the proofs of Theorems 1.1–1.3 is the fact that ridgelet analysis is intimately connected with wavelet analysis in Radon space. If $(Rf)(t,\theta)$ denotes the Radon transform of f at direction θ and position t, and τ_λ denotes the antipodally symmetrized version of $\psi^+_{j,k} \otimes w^\varepsilon_{i,\ell}$, then section 4 below shows that

$$\langle \rho_\lambda, f\rangle = \frac{1}{4\pi}\iint Rf(t,\theta)\tau_\lambda(t,\theta)dt\, d\theta. \tag{1.5}$$

Hence orthonormal ridgelet analysis amounts to a nonorthogonal wavelet analysis in Radon space. Moreover, there is a kind of Parseval relation giving an isometry between these nonorthogonal wavelet coefficients and the orthogonal ridgelet coefficients.

Because of this connection, ridgelet analysis of g is connected with wavelet analysis of Rg and, finally, the question of efficient approximation of g by ridgelets is reduced to the question of efficient approximation of Rg by wavelets. Now, if g has a singularity along a line, then Rg has a point singularity. Hence the effectiveness of ridgelet representation of objects which are smooth away from point singularities is reduced to the question of efficiency of wavelet representation of objects which are smooth away from point singularities. In sections 5 and 6 below, where the proof of Theorem 1.3 is given, this point is made by showing that the sparsity of ridgelet coefficients of objects $g(x) = g(x;\theta^0, x^0)$ is quite explicitly connected with sparsity of the high-frequency Meyer tensor wavelet coefficients of a singularity S formally defined by $S(u,v) = |v|^{-1/2}\sigma(u/|v|)$ for a certain smooth bounded function $\sigma(u)$.

1.7. Ridgelet-wavelet duality. Besides providing a useful tool in proofs, (1.5) shows an interesting duality between ridgelets and wavelets: they are good for complementary tasks.

In some sense wavelet analysis is very effective at representing objects with isolated point singularities, i.e., it takes only a few terms to obtain a reasonable approximation to such singularities. As we have just said, (1.5) means that ridgelet analysis can be very effective at representing objects with isolated point singularities *in the Radon domain*, in other words, objects with singularities along lines.

At the same time, wavelets are not efficient at representing objects with singularities along lines; nor, *therefore*, are ridgelets effective at representing objects with point singularities. Indeed, a point singularity in real space is a singularity along a sinusoidal curve in Radon space—and so, exactly as wavelets fail to deal efficiently with singularities along curves, so must ridgelets fail to deal efficiently with point singularities.

2. An orthobasis in Radon space. For a smooth function $f(x) = f(x_1, x_2)$ of rapid decay, let Rf denote the Radon transform of f, the integral along a line $\mathcal{L}_{(\theta,t)}$, expressed using the Dirac mass δ as

$$(Rf)(t,\theta) = \int f(x)\delta(x_1\cos\theta + x_2\sin\theta - t)\ dx, \tag{2.1}$$

where we permit $\theta \in [0, 2\pi)$ and $t \in \mathbb{R}$. For more information about the Radon transform see, for example, [5], [11]. Observe that the line $\mathcal{L}_{(\theta,t)}$ is identical to the line $\mathcal{L}_{(\theta+\pi,-t)}$. As a result, Rf has the *antipodal symmetry*

$$(Rf)(-t, \theta + \pi) = (Rf)(t,\theta). \tag{2.2}$$

This is a fundamental fact about the Radon transform which affects much of the notation in what follows. We adopt the convention that F (and G and variants) typically will denote a function on $\mathbb{R} \times [0, 2\pi)$ obeying the same antipodal symmetry:

$$F(-t, \theta + \pi) = F(t,\theta). \tag{2.3}$$

To create a space of such objects, we let $[\ ,\ ]$ denote the pairing

$$[F,G] = \frac{1}{4\pi}\int_0^{2\pi}\int_{-\infty}^{\infty} F(t,\theta)\bar{G}(t,\theta)dt\ d\theta, \tag{2.4}$$

and by $L^2(dt\ d\theta)$-norm we mean $\|F\|^2 = [F,F]$. Let $\mathcal{R}$ be the closed subspace of $L^2(dt\ d\theta)$ of functions F obeying (2.3). For later use, let $P_{\mathcal{R}}F$ be the orthoprojector from $L^2(dt\ d\theta)$ onto $\mathcal{R}$, defined by

$$(P_{\mathcal{R}}F)(t,\theta) = (F(t,\theta) + F(-t,\theta+\pi))/2. \tag{2.5}$$

We recall now the Meyer wavelets $\psi_{j,k}$, $j \in \mathbb{Z}$, $k \in \mathbb{Z}$, of the introduction, and the Lemarié–Meyer periodic wavelets $(w^0_{i_0,\ell} : \ell = 0, \ldots, 2^{i_0} - 1)$, $(w^1_{i,\ell} : i \geq i_0,\ \ell = 0, \ldots, 2^i - 1)$. For convenience in proofs below, we assume that the periodic Meyer wavelets are obtained by periodization of standard Meyer wavelets for $\mathbb{R}$: with a constant γ_1,

$$w^1_{i,\ell}(\theta) = \gamma_1 \cdot \sum_{h=-\infty}^{\infty} \psi_{i,\ell+h2^i}(\theta/2\pi), \quad i \geq i_0 > 0, \quad \ell = 0, \ldots, 2^i; \tag{2.6}$$

and that the periodic Lemarié scaling functions are obtained by periodization of standard Lemarié scaling functions for $\mathbb{R}$: with a constant γ_2,

$$w^0_{i_0,\ell}(\theta) = \gamma_2 \cdot \sum_{h=-\infty}^{\infty} \phi_{i_0,\ell+h2^{i_0}}(\theta/2\pi), \quad \ell = 0, \ldots, 2^{i_0}, \tag{2.7}$$

where $\phi_{i_0,\ell}$ is a standard Lemarié scaling function. However, we suppose that $\psi_{j,k}$ and $w^\varepsilon_{i,\ell}$ are normalized differently than usual, and we arrange the scaling of $\psi_{j,k}$ and the factors γ_i so

$$\|\psi_{j,k}\|_{L^2} = \sqrt{2}\ , \qquad \|w^\varepsilon_{i,\ell}\|_{L^2[0,2\pi)} = 2\sqrt{\pi}. \tag{2.8}$$

Two closure properties of these families will be important below:

$$\psi_{j,k}(-t) = \psi_{j,1-k}(t), \tag{2.9}$$

$$w^\varepsilon_{i,\ell}(\theta + \pi) = w^\varepsilon_{i,\ell+2^{i-1}}(\theta). \tag{2.10}$$

The closure property (2.9) would *not* hold for certain other prominent wavelet families, such as Daubechies' compactly supported wavelets [4].

Define the operator of reflection of functions of one variable $(Tf)(t) = f(-t)$ and the operator of translation by half a period by $(Sg)(\theta) = g(\theta+\pi)$. Note that the space $\mathcal{R}$ consists of objects invariant under $T \otimes S$; (2.3) can be rewritten $(T \otimes S)F = F$. In fact, $P_{\mathcal{R}} = (I + T \otimes S)/2$. Set now, for $j, k \in \mathbb{Z}$ and $i \geq \max(i_0, j)$, $\ell = 0, \ldots, 2^{i-1} - 1$, $\varepsilon \in \{0, 1\}$,

$$W_\lambda(t, \theta) = P_{\mathcal{R}}(\psi_{j,k} \otimes w^\varepsilon_{i,\ell}), \tag{2.11}$$

where $\lambda = (j, k; i, \ell, \varepsilon)$. For later reference, we spell this out:

$$W_\lambda(t, \theta) = (\psi_{j,k}(t) w^\varepsilon_{i,\ell}(\theta) + \psi_{j,k}(-t) w^\varepsilon_{i,\ell}(\theta + \pi))/2. \tag{2.12}$$

In a sense, the $(W_\lambda : \lambda \in \Lambda)$ make a "tensor wavelet basis with antipodal symmetry" and, like usual wavelets, their indices have a localization interpretation. j measures "ridge scale," i measures "angular scale," k measures "ridge position," ℓ measures "angular position." W_λ is localized near a *pair* of dyadic rectangles of "height" 2^{-j} and "width" $2^{-i} \cdot 2\pi$; one has lower left corner at $[t_{j,k}, \theta_{i,\ell})$ where $t_{j,k} = k/2^j$ and $\theta_{i,\ell} = \ell/2^i$; its "twin" is at $(-t_{j,k}, \theta_{i,\ell} + \pi)$.

W_λ is oscillatory in the t-direction: $\int t^m W_\lambda\ dt\ d\theta = 0\ \forall m$, owing to the oscillatory nature of Meyer wavelets. Those W_λ with $\varepsilon = 1$ are also oscillatory in the θ direction: $\int T_m(\theta) W_\lambda(t, \theta) d\theta = 0$ for each t, for each trigonometric polynomial $T_m(\theta) = \sum_{-m}^{m} c_k\ e^{ik\theta}$ of degree $m \leq m_0$, for an appropriate $m_0 = m_0(i)$; here $m_0(i) \asymp 2^i$. However, W_λ is not oscillatory in the θ-direction if $i = i_0$ and $\varepsilon = 0$: in such cases, typically $\int 1 \cdot W_\lambda(t, \theta) d\theta \neq 0$.

By construction, $W_\lambda \in \mathcal{R}$. This explains why we impose the initially unnatural-sounding restriction $\ell < 2^{i-1}$. The definition (2.11)–(2.12) also would apparently make sense for $\ell = 2^{i-1}, \dots, 2^i - 1$, but in that range, it turns out that the resulting functions W_λ are not new: indeed $W_{\lambda'} = W_\lambda$ whenever $j' = j$, $i' = i$, $k' = 1 - k$, and $\ell' = \ell + 2^{i-1}$. Our constraint $0 \le \ell < 2^{i-1}$ removes these duplications.

LEMMA 2.1. (W_λ) *is an orthobasis for* $\mathcal{R}$.

Proof. Let

$$\begin{aligned}\tilde{\Lambda} = \{(j,k;i,l,0) : j,k \in \mathbb{Z}, i = \max(i_0,j), \ell = 0,\dots,2^i-1\} \\ \cup \quad \{(j,k;i,l,1) : j,k \in \mathbb{Z}, i \ge \max(i_0,j), \ell = 0,\dots,2^i-1\};\end{aligned} \tag{2.13}$$

in comparison with Λ, note the expanded range of ℓ. The collection $(\psi_{j,k} \otimes w^\varepsilon_{i,\ell} : (j,k;i,l,\varepsilon) \in \tilde{\Lambda})$ is a complete orthonormal system for $L^2(dt\, d\theta)$. Hence $(W_\lambda)_\lambda$, which is the image of this basis under $P_{\mathcal{R}}$, is complete in $\mathcal{R}$. It remains to see that $(W_\lambda)_\lambda$ is an orthobasis. From (2.9) and (2.10), we have

$$T\psi_{j,k} = \psi_{j,1-k}, \qquad Sw^\varepsilon_{i,\ell} = w_{i,\ell+2^{i-1}}. \tag{2.14}$$

Then, for example, with $\lambda = (j,k;i,\ell,\varepsilon)$, $\lambda' = (j',k';i'\ell'\varepsilon')$,

$$[W_\lambda, W_{\lambda'}] = \frac{1}{4}\sum_{a,a'=0}^{1} [T^a\psi_{j,k} \otimes S^a w^\varepsilon_{i,\ell},\ T^{a'}\psi_{j',k'} \otimes S^{a'} w^{\varepsilon'}_{i',\ell'}]. \tag{2.15}$$

A typical term in the sum is

$$[\psi_{j,k} \otimes w^\varepsilon_{i,\ell},\ \psi_{j',k'} \otimes w^{\varepsilon'}_{i',\ell'}] = \frac{1}{4\pi}\langle \psi_{j,k}, \psi_{j',k'}\rangle (w^\varepsilon_{i,\ell},\ w^{\varepsilon'}_{i',\ell'}),$$

where here $\langle\ ,\ \rangle$ is the inner product for $L^2(dt)$ and $(\ ,\ )$ for $L^2[0,2\pi)$. Then from our normalization of $w^\varepsilon_{i,\ell}$ (see (2.8)) we have $(w^\varepsilon_{i,\ell},\ w^{\varepsilon'}_{i',\ell'}) = \delta_{ii'}\delta_{\ell\ell'}\delta_{\varepsilon\varepsilon'}4\pi$ (taking note that for $\lambda \in \Lambda$, $\varepsilon = 0$ can occur only if $i = \max(i_0,j)$), and so

$$[\psi_{j,k} \otimes w^\varepsilon_{i,\ell},\ \psi_{j',k'} \otimes w^{\varepsilon'}_{i',\ell'}] = 2 \cdot \delta_{jj'}\delta_{kk'}\delta_{ii'}\delta_{\ell\ell'}\delta_{\varepsilon\varepsilon'},$$

at least for those combinations of i,j,k,ℓ,ε and i',j',k',ℓ',ε' which can arise from indices $\lambda, \lambda' \in \Lambda$. Other terms are handled similarly, taking into account that for cross-terms

$$\begin{aligned}\langle \psi_{j,k},\ T\psi_{j',k'}\rangle &= 2 \cdot \delta_{jj'}\delta_{k,1-k'}, \\ (w^\varepsilon_{i,\ell},\ Sw^{\varepsilon'}_{i',\ell'}) &= 4\pi \cdot \delta_{ii'}\delta_{\ell\ell'+2^{i-1}}\delta_{\varepsilon\varepsilon'}.\end{aligned}$$

Since we consider only $\ell, \ell' < 2^{i-1}$, all such cross-terms vanish. It results that of the four possible combinations of terms in (2.15), only two can ever be nonzero when $\lambda, \lambda' \in \Lambda$, and so

$$[W_\lambda, W_{\lambda'}] = \frac{1}{4}\left[2 \cdot \delta_{jj'}\delta_{kk'}\delta_{ii'}\delta_{\ell\ell'}\delta_{\varepsilon\varepsilon'} + 2\delta_{jj'}\delta_{(1-k)(1-k')}\delta_{ii'}\delta_{(\ell+2^{i-1})(\ell'+2^{i-1})}\delta_{\varepsilon\varepsilon'}\right] = \delta_{\lambda,\lambda'}.$$

3. Isometry from Radon space to real space.

We now describe an isometry $\mathcal{J}$ which maps wavelets $W_\lambda \in \mathcal{R}$ to ridgelets $\rho_\lambda \in L^2(\mathbb{R}^2)$. Because the W_λ make an orthobasis for $\mathcal{R}$, the ρ_λ must make an orthobasis for a closed subspace in $L^2(\mathbb{R}^2)$. In fact the closure of the range of $\mathcal{J}$ is all of $L^2(\mathbb{R}^2)$, so that the ridgelets are a complete orthonormal system, as promised by Theorem 1.1 of the introduction. Our construction of the isometry works via the Fourier transform, and is intended to make our introductory definition (1.3) understandable. A different construction works via Radon transform ideas and will be discussed in section 4 below.

For an $F \in \mathcal{R}$, we may Fourier transform in the first variable, producing

$$\tilde{F}(\omega,\theta) = (\mathcal{F}_1 F)(\omega,\theta) = \int_{-\infty}^{\infty} F(t,\theta)\ e^{-i\omega t}\ dt,$$

where $\mathcal{F}_1$ denotes "Fourier transformation in the first variable." Let $\tilde{\mathcal{R}}$ denote the collection $\mathcal{F}_1[\mathcal{R}]$, furnished again with the inner product $[\ ,\ ]$; then

$$[\tilde{F},\tilde{G}] = 2\pi[F,G],$$

so, up to normalization, the correspondence $F \leftrightarrow \tilde{F}$ is an isometry.

For a continuous $\tilde{F} \in \tilde{\mathcal{R}}$, we may perform "polar-to-cartesian" conversion, producing a function $\hat{f}(\xi),\ \ \xi \in \mathbb{R}^2$. This works as follows: set

$$\xi(\omega,\theta) = (\omega\cos\theta,\ \omega\sin\theta), \tag{3.1}$$

where $\omega \in \mathbb{R}$ and $\theta \in [0,2\pi)$. This is a two-to-one mapping of $\mathbb{R}\times[0,2\pi)$ onto $\mathbb{R}^2$. For functions $\tilde{F} \in \tilde{\mathcal{R}}$, one can check that $\tilde{F}(\omega,\theta) = \tilde{F}(-\omega,\theta+\pi)$ and so the definition

$$\hat{f}(\xi(\omega,\theta)) = \tilde{F}(\omega,\theta)|\omega|^{-\frac{1}{2}}, \qquad (\omega,\theta) \in \mathbb{R}\times[0,2\pi), \tag{3.2}$$

is unambiguous: either of the two pairs $(\omega,\theta),(-\omega,\theta+\pi)$ giving rise to the same value of ξ provides the same definition of $\hat{f}(\xi)$. Note that if $\tilde{F}(0,\theta) \neq 0$, then the corresponding $\hat{f}(\xi)$ must have a singularity at $\xi = 0$. However, away from $\xi = 0$, $\hat{f}$ is well-defined, and in fact $\hat{f}$ is well behaved in an L^2-sense:

$$\begin{aligned}\frac{1}{2}\int_0^{2\pi}\int_{-\infty}^{\infty} |\tilde{F}(\omega,\theta)|^2\ d\omega\ d\theta &= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} |\tilde{F}(\omega,\theta)|^2\ d\omega\ d\theta \\ &= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} |\hat{f}(\xi(r,\theta))|^2\ r\ dr\ d\theta \\ &= \int |\hat{f}(\xi)|^2\ d\xi.\end{aligned}$$

Hence, because of the $\frac{1}{4\pi}$ normalizing factor in $[,]$,

$$\|\tilde{F}\|^2_{L^2(d\omega d\theta)} = \frac{1}{2\pi}\|\hat{f}\|^2_{L^2(d\xi)}. \tag{3.3}$$

We also remark that for each W_λ, the corresponding $\tilde{W}_\lambda$ is continuous and vanishes for $|\omega| < \frac{2}{3}\pi 2^j$, hence the polar-to-cartesian conversion is well-defined for every $\tilde{W}_\lambda$. More is true. Let $\hat{f} = C(\tilde{F})$ denote the operation defined by (3.2). In fact, C extends to a linear operator, well-defined on $\tilde{\mathcal{R}}$ and bounded from $\tilde{\mathcal{R}}$ to $L^2(d\xi)$ by (3.3).

In any event, the definition $\hat{\rho}_\lambda = (C\circ\mathcal{F}_1)(W_\lambda)$ makes sense, and $(\hat{\rho}_\lambda)$ is a collection of elements of $L^2(d\xi)$; the standard two-variable inverse Fourier transform $\mathcal{F}_2^{-1}$ maps this to a collection $(\rho_\lambda) \subset L^2(dx)$. The reader should now check that this definition of ρ_λ agrees with formula (1.3) in the introduction.

Put for short $\mathcal{J} = \mathcal{F}_2^{-1} \circ C \circ \mathcal{F}_1$. Evidently this is well-defined on basis elements W_λ; we now check that it is norm-preserving:

$$\begin{aligned}
\|\rho_\lambda\|^2_{L^2(dx)} &= \frac{1}{(2\pi)^2} \int |\hat{\rho}_\lambda(\xi)|^2 \, d\xi \\
&= \frac{1}{(2\pi)^2} \int_0^\pi \int_0^\infty |\hat{\rho}_\lambda(\xi(r,\theta))|^2 \, r \, dr \, d\theta \\
&= \frac{1}{(2\pi)^2} \, \frac{1}{2} \int_0^{2\pi} \int_{-\infty}^{\infty} |\tilde{W}_\lambda(\omega,\theta)|^2 |\omega|^{-1} \, |\omega| \, d\omega \, d\theta \\
&= \frac{1}{4\pi} \int_0^{2\pi} |W_\lambda(t,\theta)|^2 \, dt \, d\theta \\
&= [W_\lambda, W_\lambda] = \|W_\lambda\|_{L^2(dt \; d\theta)}.
\end{aligned}$$

In essence, we used

$$\begin{aligned}
\|\rho_\lambda\|^2_{L^2(dx)} &= \frac{1}{(2\pi)^2} \|\hat{\rho}_\lambda\|^2_{L^2(d\xi)}, \\
\|\tilde{W}_\lambda\|^2_{L^2(d\omega \; d\theta)} &= \frac{1}{2\pi} \|\hat{\rho}_\lambda\|^2_{L^2(d\xi)}, \\
\|W_\lambda\|^2_{L^2(dt \; d\theta)} &= \frac{1}{2\pi} \|\tilde{W}_\lambda\|^2_{L^2(d\omega \; d\theta)},
\end{aligned}$$

where the first and last steps are Parseval for $\mathcal{F}_2$ and $\mathcal{F}_1$ and the middle step uses (3.3).

The argument for the angle-preserving property

$$\langle \rho_\lambda, \rho_{\lambda'} \rangle = [W_\lambda, W_{\lambda'}]$$

is entirely analogous, and shows that (ρ_λ) is an orthonormal system in $L^2(\mathbb{R}^2)$.

We now show that the system is complete. Let $h \in L^1(dx) \cap L^2(dx)$ be of L^2-norm 1, and suppose that it is a bandpass function: for constants $0 < \Omega_0 < \Omega_1 < \infty$,

$$\hat{h}(\xi) = 0, \qquad |\xi| \notin [\Omega_0, \Omega_1].$$

Consider $H(t,\theta)$ defined formally by applying the adjoint of $\mathcal{J}$:

$$H = \mathcal{J}^+ h.$$

We now build up $H = \mathcal{J}^+ h$ in stepwise fashion. There is the inverse to the polar-to-cartesian transformation C, namely, the cartesian-to-polar transformation P:

$$P(\hat{h}) = \tilde{H}(\omega,\theta) = \hat{h}(\xi(\omega,\theta))|\omega|^{+\frac{1}{2}}, \qquad \omega \in \mathbb{R}, \quad \theta \in [0, 2\pi),$$

and, with one- and two-variable Fourier transforms $\mathcal{F}_1$ and $\mathcal{F}_2$ as before:

$$\mathcal{J}^+ = \mathcal{F}_1^{-1} \circ P \circ \mathcal{F}_2.$$

This mapping is well-defined on bandpass h and has an isometry property on such h:

$$\begin{aligned}
\|\hat{h}\|^2_{L^2(d\xi)} &= (2\pi)^2 \|h\|_{L^2(dx)}, \\
\|\tilde{H}\|^2_{L^2(d\omega \; d\theta)} &= \frac{1}{2\pi} \|\hat{h}\|_{L^2(d\xi)}, \\
\|H\|^2_{L^2(dt \; d\theta)} &= \frac{1}{2\pi} \|\tilde{H}\|^2_{L^2(d\omega \; d\theta)}.
\end{aligned}$$

The first and last steps are standard Parseval relations for $\mathcal{F}_2$ and $\mathcal{F}_1$, respectively. The middle step is

$$\int_0^{2\pi}\int_{-\infty}^{\infty} |\tilde{H}(\omega,\theta)|^2 \ d\omega \ d\theta = \int_0^{2\pi}\int_0^{\infty} |h(r,\theta)|^2 \ r \ dr \ d\theta$$
$$= 2\int |h(\xi)|^2 \ d\xi.$$

Hence we get $H \in \mathcal{R}$ of $L^2(dt \ d\theta)$-norm 1. One checks that

$$[H, W_\lambda] = \langle h, \rho_\lambda\rangle, \qquad \lambda \in \Lambda.$$

By hypothesis $\|h\|_{L^2(dx)} = 1$, so the sequence $([H, W_\lambda])_\lambda$ has ℓ^2-norm 1, and so the sequence $(\langle h, \rho_\lambda\rangle)_\lambda$ has ℓ^2-norm 1. It follows that there is no nontrivial unit-norm integrable bandpass function orthogonal to all the ρ_λ; as integrable bandpass functions are dense in $L^2(dx)$, the system (ρ_λ) is complete.

4. Interpretation in Radon space. We now show that analysis of f by the system (ρ_λ) is closely related to (W_λ)-wavelet analysis of the Radon transform of f. We begin by defining the adjoint of the Radon transform so that for all sufficiently nice $G \in \mathcal{R}$ and all sufficiently nice $f \in L^2(dx)$,

$$[Rf, G] = \langle f, R^+ G\rangle, \tag{4.1}$$

which leads to

$$(R^+G)(x) = \frac{1}{4\pi}\int_0^{2\pi} G(x_1\cos\theta + x_2\sin\theta, \theta) \ d\theta. \tag{4.2}$$

This operator is also called "backprojection" in the literature of computed tomography [5].

Define the Riesz order-1/2 fractional differentiation operator Δ^+ and also the order-1/2 fractional integration operator Δ^- by the unified formula

$$(\Delta^\pm f)(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\omega}\hat{f}(\omega)|\omega|^{\pm\frac{1}{2}} \ d\omega. \tag{4.3}$$

These unbounded operators are well-defined on functions which are sufficiently smooth (formally, the domain $\mathcal{D}(\Delta^+) = \{f : \int_{-\infty}^{\infty} |\hat{\rho}(\omega)|^2 \ |\omega| d\omega\}$) or sufficiently oscillatory (formally, the domain $\mathcal{D}(\Delta^-) = \{f : \int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 \ |\omega|^{-1} \ d\omega < \infty\}$); in particular, they are well-defined on every one-dimensional Meyer wavelet $\psi_{j,k}$, owing to $\operatorname{supp}(\hat{\psi}_{j,k}) \subset \{\omega : |\omega| \in [\frac{2}{3}\pi 2^j, \ \frac{8}{3}\pi 2^j]\}$. Moreover, on the appropriate domains, they are self-adjoint; and on the appropriate domains they act as inverses of each other.

Set now, for $\lambda \in \Lambda$,

$$\tau_\lambda = (\Delta^+ \otimes I)W_\lambda, \tag{4.4}$$
$$\sigma_\lambda = (\Delta^- \otimes I)W_\lambda. \tag{4.5}$$

For example,

$$\tau_{(j,k;i,\ell,1)} = (\Delta^+\psi_{j,k} \otimes w^1_{i,\ell} + \Delta^+ T\psi_{j,k} \otimes S w^1_{i,\ell})/2.$$

A useful remark is that $\Delta^\pm T = T\Delta^\pm$ on the appropriate domains.

In accord with the description of Δ^- as an "integrator" and Δ^+ as a "differentiator" we can view σ_λ as a "smoothing" of W_λ in the t-direction while τ_λ is a "roughening" of W_λ in the t-direction. This duality reflects a more fundamental biorthogonality.

Lemma 4.1. *For $\lambda \in \Lambda$,*

$$\sigma_\lambda = R\rho_\lambda, \tag{4.6}$$

$$\rho_\lambda = R^+\tau_\lambda. \tag{4.7}$$

For $\lambda, \mu \in \Lambda$,

$$[\sigma_\lambda, \tau_\mu] = \langle \rho_\lambda, \rho_\mu \rangle = [W_\lambda, W_\mu]. \tag{4.8}$$

From this we have immediately the following corollary.

Corollary 4.2. *Let f be a finite superposition of ρ_λ's. Then*

$$f = \sum_{\lambda\in\Lambda} [Rf, \tau_\lambda]\rho_\lambda, \tag{4.9}$$

$$\|f\|^2_{L^2(dx)} = \sum_{\lambda\in\Lambda} [Rf, \tau_\lambda]^2. \tag{4.10}$$

In short, the ridgelet coefficients can be read off from an analysis of the Radon transform of f, using the set τ_λ of analyzing elements. We also have, from self-adjointness of Δ^+, the following lemma.

Lemma 4.3. *Let f be a finite superposition of ρ_λ's. Then*

$$f = \sum_{\lambda\in\Lambda} [(\Delta^+ \otimes I)Rf,\ W_\lambda]\rho_\lambda, \tag{4.11}$$

$$\|f\|^2_{L^2(dx)} = \sum_{\lambda\in\Lambda} [(\Delta^+ \otimes I)Rf,\ W_\lambda]^2. \tag{4.12}$$

So the ridgelet coefficients derive from (antipodally symmetrized) wavelet analysis of the (differentiated) Radon transform.

A clarifying interpretation emerges from these representations: the justification for our use of the term "ridgelet." As introduced by [1], the term refers to a continuous ridge function

$$\psi_{a,b,\theta}(x) = a^{-\frac{1}{2}}\psi((x_1\cos\theta + x_2\sin\theta - b)/a), \tag{4.13}$$

where $\psi(t) : \mathbb{R} \to \mathbb{R}$ is oscillatory. For comparison, Lemma 4.1 gives the representation

$$\begin{aligned}\rho_\lambda(x) &= (R^+\tau_\lambda)(x)\\ &= \frac{1}{8\pi}\int_0^{2\pi} (\psi^+_{j,k}(x_1\cos\theta + x_2\sin\theta)w^\varepsilon_{i,\ell}(\theta) + \psi^+_{j,k}(-x_1\cos\theta - x_2\sin\theta)w^\varepsilon_{i,\ell}(\theta+\pi))d\theta\\ &= \frac{1}{4\pi}\int_0^{2\pi} \psi^+_{j,k}(x_1\cos\theta + x_2\sin\theta)w^\varepsilon_{i,\ell}(\theta)d\theta.\end{aligned}$$

Here we used the observation that

$$\int_0^{2\pi} \psi^+_{j,k}(x_1\cos\theta + x_2\sin\theta)w^\varepsilon_{i,\ell}(\theta)d\theta$$

$$= \int_0^\pi (\psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta) + \psi_{j,k}(x_1 \cos(\theta+\pi) + x_2 \sin(\theta+\pi)) w^\varepsilon_{i,\ell}(\theta+\pi)) d\theta$$

$$= \int_0^\pi \psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta) d\theta + \int_0^\pi \psi^+_{j,k}(-x_1 \cos\theta - x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta+\pi) d\theta.$$

Consequently,

$$\int_0^{2\pi} (\psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta) + \psi^+_{j,k}(-x_1 \cos\theta - x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta+\pi)) d\theta$$
$$= 2 \int_0^{2\pi} \psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta) w^\varepsilon_{i,\ell}(\theta) d\theta,$$

which gives Theorem 1.2 of the introduction.

Now $w^\varepsilon_{i,\ell}(\theta)$ is a function quasi-localized to an interval of length $2\pi/2^i$ near $\theta_{i,\ell} = 2\pi \cdot \ell/2^i$, with L^1-norm of size $\approx 2^{-i/2}$. Hence, roughly speaking,

$$\rho_\lambda(x) = 2^{-i/2} \text{ “Ave”}_{i,\ell}\{\psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta)\},$$

where “Ave”$_{i,\ell}$ denotes a “signed weighted average” localized near $\theta_{i,\ell}$. In short, the *ridgelet* ρ_λ is a sort of “average” of true *ridge* functions. In this “average” the weights take both positive and negative signs, which is not usual for averages; such oscillatory weights are crucial for the orthogonality properties of the ρ_λ.

We cannot expect a tighter connection of ridgelets to ridge functions than this, since ridgelets are in $L^2(dx)$ while ridge functions are not in $L^2(dx)$ owing to constancy on ridges.

To summarize the results of this section, we have the following diagram:

$$\begin{tikzcd}
 & & (\sigma_\lambda) & & \\
(\rho_\lambda) \arrow[urr, "R"] \arrow[drr, "\mathcal{F}_2"'] & & (\tau_\lambda) \arrow[ll, "R^+"'] & & (W_\lambda) \arrow[ull, "\Delta^- \otimes I"'] \arrow[ll, "\Delta^+ \otimes I"'] \arrow[d, "\mathcal{F}_1"] \\
 & & (\hat{\rho}_\lambda) & & (\tilde{W}_\lambda). \arrow[ll, "C"']
\end{tikzcd}$$

This is a commutative diagram, so that *all* routes between vertices (ρ_λ) and (W_λ) are isometries. In particular,

$$\mathcal{J} = R^+ \circ (\Delta^+ \otimes I),$$
$$\mathcal{J}^{-1} = (\Delta^+ \otimes I) \circ R.$$

In this form, the isometries $\mathcal{J}$ and $\mathcal{J}^{-1}$ are well known in the Radon transform literature; see Helgason [12, section 1.X].

5. Ridgelet analysis of a linear singularity. Recall the function $g(x_1, x_2) = 1_{\{x_2>0\}}\, e^{-x_1^2 - x_2^2}$ of the introduction. We are now in a position to show that the ridgelet coefficients of g are sparse. A more precise version of Theorem 1.3 will be proven.

THEOREM 5.1. *Let* $\alpha_\lambda = \langle \rho_\lambda, g \rangle$, $\lambda \in \Lambda$. *For each* $p \in (0, 2]$, *the ridgelet coefficients* $(\alpha_\lambda) \in \ell^p$.

The conclusion $(\alpha_\lambda) \in \ell^p$ implies that for a constant c_p,

$$\#\{\lambda : |\alpha_\lambda| \geq \delta\} \leq c_p\ \delta^{-p} \qquad \forall\ \delta > 0,$$

showing that the coefficients decay. This, in turn, implies that the partial reconstruction

$$\tilde{g}_\delta = \sum_\lambda \eta_\delta(\langle \rho_\lambda, g \rangle)\rho_\lambda$$

with $N(\delta)$-terms, $N(\delta) = \sum_\lambda 1_{\{|\alpha_\lambda| \geq \delta\}}$, obeys

$$\|g - \tilde{g}_\delta\|_{L^2} \leq C_p'\ \delta^{(1-p/2)}, \qquad N(\delta) \leq C_p\ \delta^{-p};$$

in short, $\bar{g}_N = \tilde{g}_{N(\delta)}$ gives an N-term approximation

$$\|g - \bar{g}_N\|_{L^2} \leq C_p''\ N^{-(\frac{1}{p}-\frac{1}{2})} \qquad \forall N = 1, 2, \ldots \quad \forall p \in (0, 2].$$

The proof of Theorem 5.1 shows that ridgelets are adapted to this kind of singularity *precisely* to the same extent that wavelets are adapted to point singularities; ridgelets are successful with linear singularities *because* wavelets are successful with point singularities. As section 4 showed, ridgelet analysis "is" wavelet analysis in the Radon domain, a fact we rely on heavily in this section.

5.1. Windowing and smooth change-of-variables. According to (4.11) and (4.12), the ridgelet coefficients of a function f obey

$$\alpha_\lambda = [(\Delta^+ \otimes I)(Rf), W_\lambda], \qquad \lambda \in \Lambda, \tag{5.1}$$

at least for f which are finite sums of ridge functions. It is notationally convenient for us to work using this viewpoint. For more general f, we view $(\Delta^+ \otimes I)(Rf)$ as a distribution defined to make this relation true; i.e., so that $[(\Delta^+ \otimes I)(Rf), W_\lambda] \equiv [Rf, (\Delta^+ \otimes I)W_\lambda]$. However, $(\Delta^+ \otimes I)(Rf)$ has an explicit representation as a function, provided we use this function only to calculate integrals $[(\Delta^+ \otimes I)(Rf), W_\lambda]$ against the W_λ. This point will be clarified in section 5.2 below. Since we use this function only in this way, we will be using (5.1) without any further comment.

Our choice of g was driven by the fact that we can calculate Rg. In the appendix we derive the formula

$$(Rg)(t, \theta) = e^{-t^2}\bar{\Phi}\left(-t\left|\frac{\sin\theta}{\cos\theta}\right|\right), \qquad t \in \mathbb{R}, \quad \theta \in [0, 2\pi). \tag{5.2}$$

Now Rg has point singularities of order 0—discontinuities—at points $(0, \pi/2)$ and $(0, 3\pi/2)$ (antipodal pair). Away from these points Rg is C^∞, uniformly so outside balls $B_2((0, \frac{\pi}{2}), \delta)$ and $B_2((0, \frac{3\pi}{2}), \delta)$, $\delta > 0$.

The same structural features are true of $F = (\Delta^+ \otimes I)Rg$, except that owing to the fractional derivative in the vertical direction, the point singularities of order 0 at $(0, \pi/2)$ and $(0, 3\pi/2)$ become point singularities of order $-\frac{1}{2}$. In effect, these singularities are the heart of the matter, and we now develop some windowing tools to isolate them for careful attention and a change-of-variables to move them to the origin.

Construct a C^∞ partition of unity in θ, with windows $\nu_i(\theta)$ obeying $0 \le \nu_i(\theta) \le 1$, $\sum_{i=0}^2 \nu_i(\theta) = 1$,

$$\begin{aligned}&\mathrm{supp}(\nu_1) \subset [\pi/4,\ 3\pi/4],\\ &\mathrm{supp}(\nu_2) \subset [5\pi/4,\ 7\pi/4],\\ &\mathrm{supp}(\nu_0) \subset [0, 3\pi/8] \cup \left[\frac{5\pi}{8}\ , \frac{11\pi}{8}\right] \cup \left[\frac{13\pi}{8}\ , 2\pi\right).\end{aligned}$$

In particular $\nu_1 \equiv 1$ on $[3\pi/8,\ 5\pi/8]$, while $\nu_2 \equiv 1$ on $[11\pi/8,\ 13\pi/8]$.

Define $F_i(t,\theta) = \nu_i(\theta)F(t,\theta)$; the intention is that F_1 represent behavior near $\theta = \pi/2$, F_2 represent behavior near $\theta = 3\pi/2$, and F_0 represent "everything else." Of course $F = F_0 + F_1 + F_2$. Now F_0 is C^∞ on $\mathbb{R} \times [0, 2\pi)$ and of rapid decay in t for each θ; in other words, absolutely banal. F_1 and F_2 each contain a singularity and in some sense are mirror images of each other (antipodal symmetry again); whatever goes for one will go for the other as well. We discuss F_1 only.

Introduce variables $(\bar t, \bar\theta)$ extending over all of $\mathbb{R} \times \mathbb{R}$, with $\bar t \equiv t$ and $\bar\theta \equiv \theta$ on $\theta \in \left[\frac{\pi}{4}\ , \frac{3\pi}{4}\right]$. Define the zero-extension of F_1:

$$\bar F_1(\bar t, \bar\theta) = \begin{cases} 0 & \bar\theta \notin \left[\frac{\pi}{4}\ , \frac{3\pi}{4}\right] \\ F(\bar t, \bar\theta) & \bar\theta \in \left[\frac{\pi}{4}\ , \frac{3\pi}{4}\right] \end{cases}$$

as F_1 is C^∞ at the boundary $\theta \in \{\pi/4,\ 3\pi/4\}$, this is a C^∞ extension of F_1.

Consider now the separable change-of-variables $\bar t \leftrightarrow u$, $\bar\theta \leftrightarrow v$ defined by

$$u = \bar t, \quad v = -\mathrm{cotan}(\bar\theta);$$

define this initially on subdomain $D_1 = \mathbb{R} \times \left[\frac{\pi}{4}\ , \frac{3\pi}{4}\right]$ and extend this to all $(\bar t, \bar\theta)$ while imposing four rules:

1. $u = \bar t \ \ \forall \bar\theta$;
2. separability in $(\bar t, \bar\theta) : v = v(\bar\theta)$;
3. $v(\bar\theta)$ is C^∞ in $\bar\theta$;
4. $v'(\bar\theta) = 1$ on $\bar\theta \in [\pi/8,\ 7\pi/8]^c$.

This change-of-variables induces a function

$$G_1(u,v) = \bar F_1(\bar t, \bar\theta)$$

nicely defined on all $(u,v) \in \mathbb{R}^2$; more concretely

$$G_1(u,v) = (I \otimes \bar V_1) \circ (\Delta^+ \otimes I) \circ (e^{-u^2} \otimes I)\bar\Phi(u/|v|), \tag{5.3}$$

where $\bar V_1$ is an operator of smooth windowing in v, induced from $\nu_1(\theta)$, Δ^+ is a fractional differentiation as before, and $e^{-u^2} \otimes I$ denotes a smooth windowing in u. An important point is that under this change-of-variables, the singularity at $\bar t = 0$, $\bar\theta = \pi/2$ becomes a singularity at $u = 0$, $v = 0$.

5.2. The elementary singularity. The windowing and change-of-variables now allows us to "zoom in" on the singularities in F using standard wavelet analysis. In section 5.3, we will describe a program for systematically inferring sparsity of the W_λ-coefficients from a traditional wavelet analysis of the G_i, $i = 1, 2$. However, before continuing, we digress to explain why this approach is reasonable.

Operating purely formally for the moment, the heart of the matter concerns the "elementary" singularity $S(u,v) = |v|^{-\frac{1}{2}}\sigma(u/|v|)$, where $\sigma(t) = (\Delta^+\bar{\Phi})(t)$ with Δ^+ as before and $\bar{\Phi}(t)$ as before. Indeed,

$$S(u,v) = (\Delta^+ \otimes I)\bar{\Phi}(u/|v|)$$

and so, comparing with the definition (5.3) of $G_1(u,v)$, we see that in a sense $S(u,v)$ is "what's happening" in G_1 "near $u=0,\ v=0$."

In effect, the windowing and change-of-variables have isolated our attention on the object S; if S had very *nonsparse* wavelet coefficients, then we could expect the same to be true of G_1. As we will see, S has *sparse* wavelet coefficients at fine scales.

A key remark is that S is scale-invariant—homogeneous of order $-\frac{1}{2}$. Indeed the formula $S(u,v) = |v|^{-\frac{1}{2}}\sigma(u/|v|)$ shows immediately that

$$S(au,av) = a^{-\frac{1}{2}}S(u,v). \tag{5.4}$$

It is this scale-invariance that entitles us to call S a singularity.

In order to calculate the wavelet coefficients of S, it is convenient to operate in the frequency domain, where the Meyer wavelets we are using are most naturally defined. This in turn requires a formula for "the" Fourier transform of S. However, to be clear we pause to explain that S is a special kind of distribution and that our formula for its Fourier transform works only for special purposes.

Let $\mathcal{S}(\mathbb{R})$ denote the space of Schwartz functions, functions of a single variable which are smooth and of rapid decay, along with all of their derivatives. Let $\mathcal{S}_0(\mathbb{R})$ denote the space of Schwartz functions with all moments vanishing, i.e., with $\int t^m f(t)dt = 0$ for $m = 0,1,2,\ldots$. Let $\mathcal{S}_0(\mathbb{R}) \otimes \mathcal{S}(\mathbb{R})$ denote the linear space of Schwartz functions generated from tensor products $f(u,v) = f_0(u)f_1(v)$, where $f_0 \in \mathcal{S}_0(\mathbb{R})$ and $f_1 \in \mathcal{S}(\mathbb{R})$. Such functions have classical Fourier transforms which vanish to arbitrary order on the ω_2 axis:

$$\frac{\partial^m}{\partial\omega_1^m}\frac{\partial^n}{\partial\omega_2^n}\hat{f}(\omega_1,\omega_2) = 0, \quad \omega_1 = 0, \quad m,n = 0,1,2,\ldots.$$

Such functions "omit" entirely the frequencies $\omega = (0,\omega_2)$.

We now retract our earlier definition and define S as a linear functional on $\mathcal{S}_0(\mathbb{R}) \otimes \mathcal{S}(\mathbb{R})$. Let $\dot{H}(u,v) = \bar{\Phi}(u/|v|)$; this is a bounded function of u and v and so defines a tempered distribution: $\dot{H} \in \mathcal{S}'(\mathbb{R}^2)$. Let $f \in \mathcal{S}_0(\mathbb{R}) \otimes \mathcal{S}(\mathbb{R})$; then $(\Delta^+ \otimes I)f \in \mathcal{S}(\mathbb{R}^2)$. We define S on $f \in \mathcal{S}_0 \otimes \mathcal{S}$ via

$$\langle S, f\rangle \equiv \langle \dot{H}, (\Delta^+ \otimes I)f\rangle.$$

This definition makes sense, because $\dot{H}$ is tempered and $(\Delta^+ \otimes I)f$ is in the Schwartz class.

We now give a formula for $\hat{S}$ accurate for functions in $\mathcal{S}_0 \otimes \mathcal{S}$. That is, $\langle S, f\rangle$ is correctly calculated from $\frac{1}{4\pi^2}\langle \hat{S}, \hat{f}\rangle$ when $f \in \mathcal{S}_0 \otimes \mathcal{S}$. In particular, the formula works for calculating Meyer tensor wavelet coefficients $\langle S, \psi_{j,k} \otimes \psi_{i,\ell}\rangle$, which is our only application.

LEMMA 5.2. *The singularity S, viewed as a distribution acting on $\mathcal{S}_0(\mathbb{R}) \otimes \mathcal{S}(\mathbb{R})$, has Fourier transform*

$$\hat{S}(\omega_1,\omega_2) = -2\pi\sqrt{-1}\ \mathrm{sgn}(\omega_1)|\omega_1|^{-\frac{3}{2}}\exp\{-\omega_2^2/\omega_1^2\}, \qquad \omega_1 \neq 0. \tag{5.5}$$

The proof is given in the appendix.

The earlier definition of S as a function of two variables is compatible with this definition, in the sense that when $f \in \mathcal{S}_0 \otimes \mathcal{S}$, we actually have $\langle S, f \rangle = \iint |v|^{-1/2} \sigma(u/|v|) f(u,v) du dv$, with $\sigma(u)$ the smooth bounded function defined as the inverse Fourier transform of

$$\hat{\sigma}(\omega) = \sqrt{-\pi} \cdot |\omega|^{-1/2} \cdot \mathrm{sgn}(\omega) \cdot \exp\{-\omega^2/4\}. \tag{5.6}$$

(Note that $\sigma(u)$ has rather poor decay at ∞, so it is not in L^1 or in L^2. The preceding integral nevertheless makes sense because $f \in \mathcal{S}(\mathbb{R}^2)$.)

With a formula for the Fourier transform of S in hand, we may now calculate the Meyer wavelet coefficients of S.

Lemma 5.3. *Let $\psi_{j,k}$ denote Meyer wavelets for $\mathbb{R}$, and let $\mu = (j,k,i,\ell)$ index an orthogonal tensor wavelet basis for $\mathbb{R}^2$ with elements*

$$\psi_\mu(u,v) = \psi_{j,k}(u)\psi_{i,\ell}(v), \qquad j,k,i,\ell \in \mathbb{Z}.$$

The wavelet coefficients of the elementary singularity S,

$$A_\mu = \langle S, \psi_\mu \rangle,$$

obey the exact scaling relation

$$A_{(j,k,i,\ell)} = 2^{-j/2} A_{(0,k,i-j,\ell)}. \tag{5.7}$$

For each fixed $h \geq 0$, the doubly indexed array $(A_{(0,k,h,\ell)})_{k,\ell}$ is of rapid spatial decay in k and ℓ as $|k|, |\ell| \to \infty$. In fact, for each $m > 0$,

$$|A_{(0,k,h,\ell)}| \leq C_m \ 2^{-h/2} (1+|k|)^{-m} (1+|\ell|)^{-m}, \quad k, \ell \in \mathbb{Z}. \tag{5.8}$$

As a result, the fine-scale wavelet coefficients obey

$$\sum_{j \geq i_0} \sum_{i \geq j} \sum_{k,\ell} |A_{(j,k,i,\ell)}|^p < \infty \qquad \forall \ p > 0. \tag{5.9}$$

In short, the two-dimensional tensor Meyer coefficients give a sparse representation of the high-frequency part of the elementary point singularity $S(u,v)$, with very few "big" coefficients, and with those few "big" coefficients clustered near $(k,\ell) = (0,0)$.

Proof. The scaling relation (5.4) gives, with $h = i - j$,

$$\begin{aligned} A_{jki\ell} = \langle S, \psi_{jki\ell} \rangle &= \int S(u,v) \psi_{0,k,i-j,\ell}(2^j u, 2^j v) 2^j \ du \ dv \\ &= \int S(2^j u, 2^j v) 2^{j/2} \ \psi_{0,k,h,\ell}(2^j u, 2^j v) 2^j du dv \\ &= 2^{-j/2} \langle S, \psi_{0,k,h,\ell} \rangle = 2^{-j/2} A_{0kh\ell}, \end{aligned}$$

which establishes (5.7). This scaling relation now allows us to infer (5.9) from (5.8). In fact we can see, letting $A^{(h)}$ denote the array $(A_{0kh\ell})_{k\ell}$, that (5.8) implies that $\|A^{(h)}\|_p < \infty$ for any $p > 0$; indeed we simply apply (5.8) with m chosen so that

$mp > 1$. In fact the exponential factor $2^{-h/2}$ in (5.8) yields more: we see immediately that

$$\sum_{h=0}^{\infty} \|A^{(h)}\|_p^p = A^* < \infty.$$

To see why (5.9) follows, note that

$$\begin{aligned}\sum_{j\geq i_0} \sum_{i\geq j} \sum_{k,\ell} |A_{j,k,i,\ell}|^p &= \sum_{j\geq i_0}^{\infty} \sum_{h=0}^{\infty} \sum_{k,\ell} |A_{j,k,j+h,\ell}|^p \\ &= \sum_{j\geq i_0}^{\infty} \sum_{h=0}^{\infty} \sum_{k,\ell} |2^{-j/2} A_{0,k,h,\ell}|^p \\ &= \sum_{j\geq i_0}^{\infty} \sum_{h=0}^{\infty} (2^{-j/2} \|A^{(h)}\|_p)^p \\ &= \sum_{h=0}^{\infty} \|A^{(h)}\|_p^p \sum_{j\geq i_0}^{\infty} 2^{-jp/2} = A^*/(1-2^{-i_0 p/2}).\end{aligned}$$

Hence the key point is to establish (5.8).

Working on the Fourier side, we may write

$$A_{jki\ell} = \frac{1}{4\pi^2} \int \hat{S}(\omega_1,\omega_2) \widehat{\psi_{jk}}(\omega_1) \widehat{\psi_{i\ell}}(\omega_2) d\omega_1 d\omega_2.$$

Now using $\widehat{\psi_{h\ell}}(\omega_2) = 2^{-h/2} \hat{\psi}(\omega_2/2^h) e^{-\sqrt{-1}\omega_1 \ell/2^h}$ and $\widehat{\psi_{0k}}(\omega_1) = \hat{\psi}(\omega_1) e^{-\sqrt{-1}\omega_1 k}$, we get

$$A_{0kh\ell} = \frac{1}{4\pi^2} \int \hat{S}(\omega_1,\omega_2) \hat{\psi}(\omega_1) \hat{\psi}(\omega_2/2^h) e^{-\sqrt{-1}(\omega_1 k + \omega_2 \ell/2^h)} 2^{-h/2} d\omega_1 d\omega_2.$$

Defining a new variable $\bar{\omega}_2 = \omega_2/2^h$, we get

$$A_{0kh\ell} = \frac{1}{4\pi^2} \int \hat{A}_h(\omega_1,\bar{\omega}_2) e^{-\sqrt{-1}(\omega_1 k + \bar{\omega}_2 \ell)} d\omega_1 d\bar{\omega}_2,$$

where we defined

$$\hat{A}_h(\omega_1,\bar{\omega}_2) = \hat{S}(\omega_1,\bar{\omega}_2 2^h) \hat{\psi}(\omega_1) \hat{\psi}(\bar{\omega}_2) 2^{h/2}.$$

In short,

$$A_{0kh\ell} = A_h(k,\ell), \qquad k,\ell \in \mathbb{Z},$$

where

$$A_h(u,v) = \frac{1}{(2\pi)^2} \int \hat{A}_h(\omega_1,\omega_2)\ e^{-\sqrt{-1}(\omega_1 u + \omega_2 v)} d\omega_1 d\omega_2.$$

We note that $\hat{A}_h$ is supported in $\Omega = \{(\omega_1,\omega_2) : |\omega_i| \in [\frac{2\pi}{3}\ , \frac{8\pi}{3}], i = 1,2\}$, and that $\hat{A}_h \in C_0^\infty[\Omega]$. Now from $|f(u,v)| \leq \int |\hat{f}(\omega_1,\omega_2)| d\omega_1\ d\omega_2$ and from

$$(-\sqrt{-1})^{m+n} \cdot u^m v^n \cdot f(u,v) = \frac{1}{(2\pi)^2} \int \frac{\partial^m}{\partial^m \omega_1} \frac{\partial^n}{\partial^n \omega_2} \hat{f}(\omega_1,\omega_2) e^{-\sqrt{-1}(u\omega_1 + v\omega_2)} d\omega_1 d\omega_2$$

valid for $m, n = 0, 1, 2, \ldots$ and all sufficiently nice f, we get that for $m = 1, 2, \ldots$,

$$|A_h(k,\ell)| \le C_m \cdot \|\hat{A}_h\|_{W_1^m[\mathbb{R}^2]}(1+|k|)^{-m}(1+|\ell|)^{-m}, \quad k, \ell \in \mathbb{Z},$$

where the $W_1^m[\mathbb{R}^2]$-norm of a smooth function f is simply $\|f\|_{L^1[\mathbb{R}^2]} + \|f^{(m)}\|_{L^1[\mathbb{R}^2]}$. Consequently, (5.8) reduces merely to the assertion that for each $m = 1, 2, 3, \ldots$, there is C_m so that

$$\|\hat{A}_h\|_{W_1^m[\mathbb{R}^2]} \le C_m 2^{-h/2}.$$

Now write

$$\hat{A}_h(\omega_1,\omega_2) = B_h(\omega_1,\omega_2) \cdot \Psi(\omega_1,\omega_2), \tag{5.10}$$

where

$$\begin{aligned} \Psi(\omega_1,\omega_2) &= -2\pi\ \sqrt{-1}\ \mathrm{sgn}(\omega_1)\ |\omega_1|^{-\frac{3}{2}}\ \hat{\psi}(\omega_1)\hat{\psi}(\omega_2), \\ B_h(\omega_1,\omega_2) &= 2^{h/2}\exp\{-2^{2h}\omega_2^2/\omega_1^2\}. \end{aligned}$$

Now for a function $f \in C_0^m[\Omega]$, $\|f\|_{W_1^m[\mathbb{R}^2]} \le C'_m \|f\|_{C^m[\Omega]}$, where the $C^m[\Omega]$-norm of f refers to the smallest constant C bounding the $L^\infty[\Omega]$-norm of f and also the $L^\infty[\Omega]$ norm of every mixed partial of total degree $\le m$. Using the fact that Ψ and all of its derivatives vanish on $\partial\Omega$,

$$\|\hat{A}_h\|_{C^m[\Omega]} \le C''_m \|B_h\|_{C^m[\Omega]} \cdot \|\Psi\|_{C^m[\Omega]}, \tag{5.11}$$

and, since $\|\Psi\|_{C^m[\Omega]} \le C'''_m$, $m = 1, 2, \ldots$, we conclude that for each $m = 1, 2, \ldots$ we have an inequality of the form

$$\|\hat{A}_h\|_{W_1^m[\mathbb{R}^2]} \le C_m^{(iv)} \|B_h\|_{C^m[\Omega]}.$$

The proof is thus finished by the following lemma, which gives the required bounds on the norms of B_h.

Lemma 5.4. *For $h \ge 0$, and all $\lambda > 0$,*

$$\|B_h\|_{C^m[\Omega]} \le C_{\lambda,m}\ 2^{-\lambda 2^h}, \qquad m = 1, 2, \ldots. \tag{5.12}$$

The lemma is proved in the appendix.

5.3. Program of proof. Now that we know that the singularities "at the heart" of F have in a certain sense sparse coefficients, we are encouraged to elaborate the approach into a systematic proof that F itself has sparse coefficients.

The change-of-variables and windowing operations of section 5.1 created a series of objects which may be related to our object of interest F by

$$F = \sum_{r=1}^{2}(\tau_{1,r} \circ \tau_{2,r} \circ \tau_{3,r})G_r + (\tau_{1,0} \circ \tau_{2,0})\bar{F}_0. \tag{5.13}$$

Here the $\tau_{j,r}$ are change-of-variables operations and $\bar{F}_0$, G_1, and G_2 were all defined in section 5.1.

The terms in this sum may be arranged as in Table 5.1. In this table, cells in the same column describe objects defined on a common domain, and adjacent cells in the

Table 5.1
Relations among various functions.

	Col. 0		1		2		3
Row 0		$\overset{\tau_{1,0}}{\swarrow}$	$F_0(t,\theta)$	$\xleftarrow{\tau_{2,0}}$	$\bar{F}_0(\bar{t},\bar{\theta})$		
1	$F=\sum$	$\xleftarrow{\tau_{1,1}}$	$F_1(t,\theta)$	$\xleftarrow{\tau_{2,1}}$	$\bar{F}_1(\bar{t},\bar{\theta})$	$\xleftarrow{\tau_{3,1}}$	$G_1(u,v)$
2		$\overset{\tau_{1,2}}{\nwarrow}$	$F_2(t,\theta)$	$\xleftarrow{\tau_{2,2}}$	$\bar{F}_2(\bar{t},\bar{\theta})$	$\xleftarrow{\tau_{3,2}}$	$G_2(u,v)$

same row are objects on different domains linked by a transformation. For example $F_1(t,\theta) \xleftarrow{\tau_{2,1}} \bar{F}_1(\bar{t},\bar{\theta})$ means that there is a transformation $\tau_{2,1}$ such that $F_1 = \tau_{2,1}\bar{F}_1$; this is the change-of-variables transformation transporting a function from domain $(\bar{t},\bar{\theta})$ to the "same" function on domain (t,θ). Corresponding to each column in the table is a domain. For example, in Column 3, we have the domain of all $(u,v) \in \mathbb{R}^2$, while in Column 2, we have the domain of $(\bar{t},\bar{\theta}) \in \mathbb{R}^2$. We may associate to each such domain an orthobasis. For the domain of (u,v), we associate the orthobasis ψ_μ, where $\mu = (j,k,i,l,\varepsilon)$, and where μ runs through the set

$$\begin{aligned} M = \{&(j,k,i,\ell,0) : j,k,\ell \in \mathbb{Z}, i = \max(i_0,j)\} \\ &\cup\{(j,k,i,\ell,1) : j,k,\ell \in \mathbb{Z},\ i \geq \max(i_0,j)\}. \end{aligned} \tag{5.14}$$

For the domain of $(\bar{t},\bar{\theta})$, we associate the orthobasis $\bar{\psi}_{\bar{\mu}}(\bar{t},\bar{\theta})$, where $\bar{\mu}$ runs through the set M. For the domain of $(t,\theta) \in \mathbb{R} \times [0,2\pi)$, which occurs both in Columns 0 and 1, we associate two bases. For Column 0 we associate the antipodally symmetrized basis (W_λ) of section 2. For Column 1 we associate a standard tensor basis. Table 5.2 summarizes these choices. We may expand an object in a given column of Table 5.1 in the corresponding orthobasis for that column. We denote coefficients in Column 3 by A^1 and A^2, in Column 2 by B^0, B^1, and B^2, and coefficients in Column 1 by C^0, C^1, and C^2. Thus, for example,

$$\begin{aligned} G_1 &= \sum_\mu A^1_\mu\ \psi_\mu, \\ \bar{F}_1 &= \sum_{\bar{\mu}} B^1_{\bar{\mu}}\ \bar{\psi}_{\bar{\mu}}, \\ F_1 &= \sum_{\tilde{\lambda}} C^1_{\tilde{\lambda}}\ \tilde{\psi}_{\tilde{\lambda}}, \\ F &= \sum_\lambda \alpha_\lambda\ W_\lambda. \end{aligned}$$

Now objects in the same row of the Columns 2 and 3 of Table 5.1 are just "the same" object in different coordinate systems (i.e., we have $\bar{F}_1 = \tau_{3,1}G_1$, etc.), so the coefficients (A^r_μ) and $(B^r_{\bar{\mu}})$ for the same value of r are linearly related. Indeed, let $\psi^*_\mu(\bar{t},\bar{\theta})$ denote the "pushout" of ψ_μ from (u,v) coordinates to $(\bar{t},\bar{\theta})$ coordinates:

$$\psi^*_\mu(\bar{t},\bar{\theta}) = \psi_\mu(u(\bar{t}),v(\bar{\theta}));$$

then for $r = 1,2$ we have matrices $T_{3,r}$ such that

$$B^r = T_{3,r}A^r, \qquad r = 1,2,$$

Table 5.2
Relations between domains.

	Col. 0	1	2	3
Domain	(t,θ)	(t,θ)	$(\bar{t},\bar{\theta})$	(u,v)
Basis element	$W_\lambda(t,\theta)$	$\tilde{\psi}_{\tilde{\lambda}}(t,\theta)$	$\bar{\psi}_{\bar{\mu}}(\bar{t},\bar{\theta})$	$\psi_\mu(u,v)$
Formula	(2.12)	$\psi_{j,k}(t)w^{\varepsilon}_{i,\ell}(\theta)$	$\psi_{j,k}(\bar{t})\psi^{\varepsilon}_{i,\ell}(\theta)$	$\psi_{j,k}(u)\psi^{\varepsilon}_{i,\ell}(v)$
Index	λ	$\tilde{\lambda}=(j,k,i,l,\varepsilon)$	$\bar{\mu}=(j,k,i,l,\varepsilon)$	$\mu=(j,k,i,l,\varepsilon)$
Index set	Λ	$\{\tilde{\lambda}\}=\tilde{\Lambda}$	$\{\bar{\mu}\}=M$	$\{\mu\}=M$

where

$$(T_{3,r})_{\bar{\mu},\mu} = \langle \bar{\psi}_{\bar{\mu}}, \psi^*_\mu \rangle_{L^2(d\bar{t}d\bar{\theta})}.$$

Viewed another way, we have the diagram

$$\begin{tikzcd}
\bar{F}_1 & G_1 \arrow[l, "\tau_{3,1}"'] \arrow[d, "\mathcal{W}"] \\
B^1 \arrow[u, "\overline{\mathcal{W}}^{-1}"] & A^1, \arrow[l, "T_{3,1}"']
\end{tikzcd}$$

where $\mathcal{W}$ denotes wavelet transform using the ψ_μ basis, and $\overline{\mathcal{W}}^{-1}$ denotes inverse wavelet transform using the $\bar{\psi}_{\bar{\mu}}$ basis. That is, $T_{3,r}$ is the matrix representation of transformation $\tau_{3,r}$.

Also, as the $(\bar{t},\bar{\theta})$ and (t,θ) coordinate systems are related by periodization, and the objects in Columns 1 and 2 of the sa me row of Table 5.1 are identical on the common domain, the coefficients (B^r_μ) and $(C^r_{\tilde{\lambda}})$ are linearly related. Indeed, from (2.6) we have, for $\tilde{\lambda}=(j,k,i,\ell,\varepsilon)$,

$$C^r_{(j,k,i,\ell,\varepsilon)} = \sum_{h=-\infty}^{\infty} B^r_{(j,k,i,\ell+h2^i,\varepsilon)};$$

hence there is a matrix $T_{2,r}$ with

$$C^r = T_{2,r}B^r.$$

Finally, as

$$W_{(j,k,i,\ell,\varepsilon)} = (\tilde{\psi}_{(j,k,i,\ell,\varepsilon)} + \tilde{\psi}_{(j,1-k,i,\ell+2^{i-1},\varepsilon)})/2,$$

we have

$$2[W_{(j,k,i,\ell,\varepsilon)}, F] = [\tilde{\psi}_{(j,k,i,\ell,\varepsilon)}, F] + [\tilde{\psi}_{(j,1-k,i,\ell+2^{i-1},\varepsilon)}, F],$$

and so there is a matrix T_1 with

$$\alpha = T_1C^1 + T_1C^2 + T_1C^0.$$

That is, the same matrix T_1 represents all three transformations $\tau_{1,r}$.

Combining these remarks,

$$\alpha = \sum_{r=1}^{2}(T_1 \circ T_{2,r} \circ T_{3,r})A^r + (T_1 \circ T_{2,0})B^0. \tag{5.15}$$

Table 5.3
Relations between coefficient arrays.

	Col. 0		1		2		3
Row 0		$\overset{T_1}{\swarrow}$	C^0	$\overset{T_{2,0}}{\longleftarrow}$	B^0		
1	$\alpha = \sum$	$\overset{T_1}{\longleftarrow}$	C^1	$\overset{T_{2,1}}{\longleftarrow}$	B^1	$\overset{T_{3,1}}{\longleftarrow}$	A^1
2		$\overset{T_1}{\nwarrow}$	C^2	$\overset{T_{2,2}}{\longleftarrow}$	B^2	$\overset{T_{3,2}}{\longleftarrow}$	A^2

In short, we have a tabular arrangement at the level of coefficients as summarized in Table 5.3. Suppose that we can show that the A^r are in every ℓ^p for $r = 1, 2$, and that B^0 is in every ℓ^p, and also that every T-mapping is bounded from ℓ^p to ℓ^p for every p, $0 < p \leq 1$. It will then follow that $\alpha \in \ell^p$ for every $p > 0$.

Theorem 5.1 therefore follows from these lemmas, which focus on the range $0 < p \leq 1$, which is the important one.

Lemma 5.5. *For each $p \in (0, 1]$, T_1 is a bounded linear mapping from $\ell^p(\tilde{\Lambda})$ to $\ell^p(\Lambda)$.*

Lemma 5.6. *For each $p \in (0, 1]$, and for $r = 0, 1, 2$, $T_{2,r}$ is a bounded linear mapping from $\ell^p(M)$ to $\ell^p(\tilde{\Lambda})$.*

Lemma 5.7. *For each $p \in (0, 1]$, and for $r = 1, 2$, $T_{3,r}$ is a bounded linear mapping from $\ell^p(M)$ to $\ell^p(M)$.*

Lemma 5.8. *For each $p \in (0, 1]$, for $r = 1, 2$, $A^r \in \ell^p(M)$.*

Lemma 5.9. *For each $p \in (0, 1]$, $B^0 \in \ell^p(M)$.*

The "hardest" of these lemmas is Lemma 5.8; the analysis of section 5.2 suggests why it should be true; in effect G_1 behaves like S at high resolutions, and S has its high-resolution wavelet coefficients in every ℓ^p.

5.4. Easy pieces.

We begin with the easy Lemmas 5.5, 5.6, and 5.9.

Lemma 5.5, concerning T_1, is just the p-triangle inequality: for $p \leq 1$,

$$|C^r_{(j,k,i,\ell,\varepsilon)} + C^r_{(j,1-k,i,\ell+2^{i-1},\varepsilon)}|^p \leq |C^r_{(j,k,i,\ell,\varepsilon)}|^p + |C^r_{(j,1-k,i,\ell+2^{i-1},\varepsilon)}|^p;$$

summing over $\lambda \in \Lambda$ we get

$$\sum_{\lambda \in \Lambda} |C^r_{(j,k,i,\ell,\varepsilon)} + C^r_{(j,1-k,i,\ell+2^{i-1},\varepsilon)}|^p \leq \sum_{\tilde{\lambda} \in \tilde{\Lambda}} |C^r_{\tilde{\lambda}}|^p.$$

Lemma 5.6, concerning $T_{2,r}$, is also the p-triangle inequality:

$$\left| \sum_{h=-\infty}^{\infty} B^r_{(j,k,i,\ell+2^i \cdot h,\varepsilon)} \right|^p \leq \sum_{h=-\infty}^{\infty} |B^r_{(j,k,i,\ell+h2^i,\varepsilon)}|^p,$$

from which follows

$$\|C^r\|_{\ell^p} \leq \|B^r\|_{\ell^p}, \qquad r = 0, 1, 2; \quad p \in (0, 1].$$

Lemma 5.9, concerning B^0, is just an observation: if $\bar{F}_0(\bar{t}, \bar{\theta})$ is C^∞, with $\bar{F}_0$ and all its partial derivatives of rapid decay as $t \to \infty$, then its wavelet coefficients are in every ℓ^p, $p \in (0, 2]$.

5.5. Boundedness of $T_{3,r}$. We now consider Lemma 5.7. In principle, this is a simple matter, about the well-behavedness of the tensor wavelet transform under separable changes-of-variables. However, some of the estimation ideas play an important role in section 6, so we spell them out carefully.

Put for short $\mathbf{t}_{\mu,\mu'} = \langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle$. The norm of $T_{3,r}$ is bounded by

$$\|T_{3,r}\|_{(\ell^p \to \ell^p)} \le \sup_{\mu'} \left(\sum_\mu |\mathbf{t}_{\mu,\mu'}|^p \right)^{1/p}.$$

Let $M_0 = \{\mu : \varepsilon = 0\}$, $M_1 = \{\mu : \varepsilon = 1\}$, and let, for a,$b \in \{0,1\}$,

$$S_{a,b} = \sup_{\mu' \in M_a} \sum_{\mu \in M_b} |\mathbf{t}_{\mu,\mu'}|^p.$$

We need to show that

$$S_{a,b} < \infty, \qquad a, b \in \{0,1\}.$$

We first remark that $t_{\mu,\mu'}$ can be nonzero only if $j = j'$ and $k = k'$. We now consider four cases:

$\underline{S_{0,0}}$: here $i = i' = \max(i_0, j)$, $\varepsilon = \varepsilon' = 0$. By rapid decay of $\bar{\psi}_\mu$ and $\psi^*_{\mu'}$, we get for $m = 1, 2, \ldots$,

$$|\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle| \le C_m \; 2^{i/2} (1 + 2^i |\bar{\theta}_{i,\ell} - \bar{\theta}^*_{i',\ell'}|)^{-m} \cdot \delta_{jj'} \delta_{kk'}.$$

Here $\bar{\theta}_{i,\ell} = \ell/2^i \cdot 2\pi$, $\bar{\theta}^*_{i',\ell'} = \bar{\theta}(\ell'/2^{i'})$. Picking m so large that $mp > 1$, and setting $t' = 2^{i'} \bar{\theta}^*_{i',\ell'}$

$$\sum_{\mu \in M_0} |\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle|^p \le C_m \; 2^{ip/2} \sum_\ell (1 + |\ell - t'|)^{-mp} \le C_{m,p} < \infty.$$

$\underline{S_{0,1}}$: here $\varepsilon' = 0$, $\varepsilon = 1$. Now $\psi^*_{\mu'}$ is a smooth function at scale $2^{-i'}$, and as $i \ge \max(i_0, j) = \max(i_0, j') = i'$, $\bar{\psi}_\mu$ is an oscillatory function at a finer scale. As $\bar{\psi}_\mu$ has more than m vanishing moments, the m-fold integration by parts formula

$$\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle = (-1)^m \langle \bar{\psi}_\mu^{(-m)}, (\psi^*_{\mu'})^{(m)} \rangle,$$

gives, for each $m = 1, 2, 3, \ldots$,

$$|\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle| \le C_m \; 2^{-(i-i')(m-\frac{1}{2})} (1 + 2^{i'} |\bar{\theta}_{i,\ell} - \bar{\theta}^*_{i',\ell'}|)^{-m} \cdot \delta_{jj'} \delta_{kk'}.$$

Pick $m > 1/p + 1/2$. With $h = i - i'$ and $t' = 2^{i'} \bar{\theta}^*_{i',\ell'}$,

$$\begin{aligned} \sum_{\mu \in M_1} |\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle|^p &\le C \sum_{h \ge 0} 2^{-h(m-\frac{1}{2})p} \sum_\ell (1 + |2^{-h}\ell - t'|)^{-mp} \\ &\le C \sum_{h \ge 0} 2^{-h(m-\frac{1}{2})p} 2^h. \end{aligned}$$

As $(m - \frac{1}{2})p > 1$, we get $S_{0,1} < \infty$.

$\underline{S_{1,0}}$: here $\varepsilon = 0$, $\varepsilon' = 1$. Now we reverse viewpoint. $\bar{\psi}_\mu$ is a smooth function at scale 2^{-i}, $i' \ge \max(i_0, j') = \max(i_0, j) = i$, and $\psi_{\mu'}$ is an oscillatory function at fine scale. Writing $\bar{\psi}^*_\mu$ for the "pullback" $\bar{\psi}_\mu(\bar{t}(u), \bar{\theta}(v))$, and $J(u,v)$ for the Jacobian $\left|\frac{d\bar{\theta}}{dv}\right|$, then

$$\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle = \langle \bar{\psi}^*_\mu \cdot J, \psi_{\mu'} \rangle.$$

Now $\bar{\psi}_\mu \cdot J$ is smooth and of rapid decay, together with all its partial derivatives, so we get for $m = 1, 2, 3, \ldots$

$$|\langle \bar{\psi}^*_\mu \cdot J, \psi_{\mu'} \rangle| \le C_m \; 2^{-(i'-i)(m-\frac{1}{2})} (1 + 2^i |v^*_{i,\ell} - v_{i',\ell'}|)^{-m} \cdot \delta_{jj'} \delta_{kk'},$$

where $v_{i',\ell'} = \ell'/2^i$ and $v^*_{i,\ell} = v(\bar{\theta}_{i,\ell})$. Pick $m > 1/p$. Put $\ell^*(\ell) = 2^i v^*_{i,\ell}$. Now

$$|\ell^*(\ell) - \ell^*(\ell')| \ge c|\ell - \ell'|,$$

where $c = \inf_{\bar{\theta}} |v'(\bar{\theta})| > 0$; whatever t' may be,

$$\sum_\ell (1 + |\ell^*(\ell) - t'|)^{-mp} < C_{m,p}.$$

It follows that

$$\begin{aligned} \sum_{\mu \in M_0} |\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle|^p &\le C \cdot \sum_{h \ge 0} 2^{-h(m-\frac{1}{2})p} \sum_\ell (1 + |\ell^*(\ell) - 2^{-h}\ell'|)^{-mp} \\ &\le C \cdot \sum_{h \ge 0} 2^{-h(m-\frac{1}{2})p}. \end{aligned}$$

As $(m - \frac{1}{2})p > 0$, we get $S_{1,0} < \infty$.

$\underline{S_{1,1}}$: here we have $i, i' \ge \max(i_0, j)$, $\varepsilon = \varepsilon' = 1$. We consider both cases where $\bar{\psi}_\mu$ is viewed as the oscillatory member and, alternatively, as the smooth member of the pair. In the case $i' \ge i$, $\bar{\psi}_\mu$ is the smooth member of the pair; arguing as in case $S_{1,0}$ gives

$$|\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle| \le C \cdot 2^{-(i'-i)(m-\frac{1}{2})} (1 + 2^i |v^*_{i,\ell} - v_{i',\ell'}|)^{-m} \cdot \delta_{jj'} \delta_{kk'},$$

while if $i' \le i$, $\bar{\psi}_\mu$ is the oscillatory member of the pair; arguing as in case $S_{0,1}$ gives

$$|\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle| \le C \cdot 2^{-(i-i')(m-\frac{1}{2})} (1 + 2^{i'} |\bar{\theta}_{i,\ell} - \bar{\theta}^*_{i',\ell'}|)^{-m} \cdot \delta_{jj'} \delta_{kk'}.$$

Pick $m > 1/2 + 1/p$. Using the above cases, and arguing as earlier, we get

$$\begin{aligned} \sum_{\mu \in M_1} |\langle \bar{\psi}_\mu, \psi^*_{\mu'} \rangle|^p &\le C \sum_{i < i'} \sum_\ell + \sum_{i \ge i'} \sum_\ell \\ &\le C \sum_{h \ge 0} 2^{-h(m-\frac{1}{2})p} + C \sum_{h \ge 0} 2^{-h(m-\frac{1}{2})p} \; 2^h. \end{aligned}$$

Because $(m - \frac{1}{2})p > 1$, we get $S_{1,1} < \infty$.

5.6. Sparsity of coefficients (A^r_μ). Now we arrive at the "heart of the matter" a second time, this time "for real." In essence, we refine the argument of Lemma 5.3. The key point is that Lemma 5.3 studied a precisely scale-invariant object, whereas now we study an object which is only asymptotically scale-invariant.

The argument is the same for $r = 1, 2$, so we consider only $r = 1$ and we omit the superscript 1.

We define $A_\mu = \langle H, (\Delta^+ \otimes I)\psi_\mu\rangle$, where

$$H(u, v) = \bar{\Phi}(u/|v|)V(u, v),$$

with $V(u, v) = e^{-u^2} \cdot \bar{V}_1(v)$ a renaming of the windowing terms in (5.3).

First, we consider the case $\varepsilon = 0$ and show that

$$\sum_{j=-\infty}^{\infty} \sum_{i=\max(i_0,j)} \sum_{k,\ell=-\infty}^{\infty} |A_{jki\ell 0}|^p \leq C^p.$$

In case $\varepsilon = 0$ and $j \leq i_0$, the sparsity of coefficients A_μ follows just from the rapid decay of $V(u, v)$ and boundedness of $\bar{\Phi}(u/|v|)$, which give

$$\sum_{j=-\infty}^{i_0} \sum_{i=i_0} \sum_{k,\ell=-\infty}^{\infty} |A_{jki\ell 0}|^p \leq C^p.$$

In case $\varepsilon = 0$ and $j > i_0$, the sparsity of coefficients A_μ follows from the argument used below in the case $\varepsilon = 1$; indeed it may be seen that the argument used there for $\Omega = \{\omega : |\omega_i| \in [2\pi/3, 8\pi/3]\}$ adapts easily to the larger set $\Omega' = \{\omega : |\omega_1| \in [2\pi/3, 8\pi/3], \omega_2 \in [-8\pi/3, 8\pi/3]\}$. This adaptation will yield

$$\sum_{j\geq i_0} \sum_{i=j} \sum_{k,\ell=-\infty}^{\infty} |A_{jki\ell 0}|^p \leq C^p.$$

So consider now the case $\varepsilon = 1$. To begin, we need a formula for the Fourier transform of $H(u, v)$, viewed as a tempered distribution in $\mathcal{S}'(\mathbb{R}^2)$. The convolution formula for tempered distributions [15] says that if $f \in \mathcal{S}'(\mathbb{R}^2)$ and $g \in \mathcal{S}(\mathbb{R}^2)$, then $(f \cdot g)\hat{} = (2\pi)^{-2}\hat{f} \star \hat{g}$. Hence

$$\hat{H} = (2\pi)^{-2} \cdot (\widehat{\dot{H}} \star \hat{V}) = \gamma_1\hat{V} + \gamma_2\hat{V} \star \beta,$$

where $\dot{H}$ is as in Lemma 8.1, and where the constants γ_r, $r = 1, 2$ can be obtained from that lemma, and where

$$\beta(\omega_1, \omega_2) = \mathrm{sgn}(\omega_1) \cdot |\omega_1|^{-2} \cdot e^{-\omega_2^2/\omega_1^2}.$$

Now obviously

$$\begin{aligned} A_\mu &= \gamma_1\langle \hat{V}, ((\Delta^+ \otimes I)\psi_\mu)\hat{}\rangle + \gamma_2\langle \hat{V} \star \beta, ((\Delta^+ \otimes I)\psi_\mu)\hat{}\rangle \\ &= \gamma_1\bar{A}_\mu + \gamma_2\tilde{A}_\mu, \end{aligned}$$

say. The first term in this expression has

$$\sum_{j=-\infty}^{\infty} \sum_{i\geq\max(i_0,j)} \sum_{k,\ell=-\infty}^{\infty} |\bar{A}_{jki\ell 1}|^p \leq C^p,$$

because V and all of its derivatives are smooth and of rapid decay. So we turn to the second term in the expression, ignoring the constant factor γ_2. To review, we wish to establish

$$\sum_{j=-\infty}^{\infty} \sum_{i\geq\max(i_0,j)} \sum_{k,\ell=-\infty}^{\infty} |\tilde{A}_{jki\ell 1}|^p \leq C^p, \tag{5.16}$$

where

$$\tilde{A}_{jki\ell 1} = \langle \hat{V} \star \beta, ((\Delta^+ \otimes I)\psi_\mu)\hat{\ } \rangle.$$

Now define $\beta_0(\xi_1, \xi_2) = (\hat{V} \star \beta)(\xi_1, \xi_2)$, and $\beta_j(\xi_1, \xi_2) = (V_j \star \beta)(\xi_1, \xi_2)$, where

$$V_j(\xi) = 2^{2j} \hat{V}(2^j \xi_1, 2^j \xi_2);$$

note that we are omitting the hat,$\hat{\ }$, from V_j, despite that fact that V_j is acting on the Fourier side and is defined in terms of $\hat{V}$. Below we will use the scaling relation

$$\beta_0(2^j\xi_1, 2^j\xi_2) = 2^{-2j}\beta_j(\xi_1, \xi_2). \tag{5.17}$$

Now

$$\begin{aligned}
\tilde{A}_{jki\ell\varepsilon} &= \iint \beta_0(\omega_1, \omega_2)|\omega_1|^{1/2}\hat{\psi}_{jk}(\omega_1)\hat{\psi}_{i\ell}(\omega_2)d\omega_1 d\omega_2 \\
&= \iint \beta_0(\omega_1, \omega_2)|\omega_1|^{1/2}\hat{\psi}(\omega_1/2^j)\hat{\psi}(\omega_2/2^i) \\
&\qquad \cdot \exp\{-\sqrt{-1}(\omega_1 k/2^j + \omega_2 \ell/2^i)\}2^{-(i+j)/2}d\omega_1 d\omega_2 \\
&= \iint \beta_0(\xi_1 2^j, \xi_2 2^i)|\xi_1 2^j|^{1/2}\hat{\psi}(\xi_1)\hat{\psi}(\xi_2) \\
&\qquad \cdot \exp\{-\sqrt{-1}(\xi_1 k + \xi_2 \ell)\}2^{(i+j)/2}d\xi_1 d\xi_2 \\
&= \iint_\Omega \hat{A}_{j,h}(\xi_1, \xi_2)\exp\{-\sqrt{-1}(\xi_1 k + \xi_2 \ell)\}d\xi_1 d\xi_2,
\end{aligned}$$

say, where we made the change-of-variables $\xi_1 = \omega_1/2^j$, $\xi_2 = \omega_2/2^i$, used the scaling relation (5.17), and where once again $\Omega = \{\omega : |\omega_i| \in [\frac{2\pi}{3}, \frac{8\pi}{3}]\}$. Here

$$\begin{aligned}
\hat{A}_{j,h}(\xi_1, \xi_2) &= \beta_j(\xi_1, \xi_2 2^h)|\xi_1|^{1/2}\hat{\psi}(\xi_1)\hat{\psi}(\xi_2)2^{-2j}2^{i/2}2^j \\
&= B_{j,h}(\xi_1, \xi_2) \cdot \Psi(\xi_1, \xi_2),
\end{aligned}$$

say, where

$$B_{j,h}(\xi_1, \xi_2) = \beta_j(\xi_1, \xi_2 2^h)2^{h/2}2^{-j/2}, \tag{5.18}$$

$$\Psi(\xi_1, \xi_2) = |\xi_1|^{1/2} \cdot \hat{\psi}(\xi_1)\hat{\psi}(\xi_2).$$

By the same type of analysis as in Lemma 5.3, we can conclude (5.16) once we establish that, for $m = 1, 2, 3, \ldots$, we have

$$\sum_{j \geq i_0} \sum_{h=0}^{\infty} \|\hat{A}_{j,h}\|^p_{C^m[\Omega]} < \infty \quad \forall p > 0.$$

Also as in (5.11) in that lemma, this will follow from the C^∞ nature of the compactly supported function $\Psi(\xi_1, \xi_2)$ and the following estimate on $B_{j,h}$.

LEMMA 5.10. *$B_{j,h} \in C^\infty[\Omega]$ $\forall$ $j \geq i_0$, and $h \geq 0$. In addition,*

$$\sum_{j \geq i_0} \sum_{h \geq 0} \|B_{j,h}\|^p_{C^m[\Omega]} < \infty, \qquad m = 1, 2, 3, \ldots. \tag{5.19}$$

The lemma is proved in the appendix.

6. Analysis of a more general linear singularity. We now consider the more general mutilated Gaussian

$$g_{\theta^0,x^0}(x) = g^0(U_{\theta^0}(x - x^0)),$$

where now g^0 is the "standard mutilated Gaussian" that was called g in section 5, U_{θ^0} is rotation by θ^0, and x^0 is a choice of origin. We will see that the sparsity properties of the ridgelet coefficients of g^0 hold equally for the ridgelet coefficients of g_{θ^0,x^0}.

The key point is to invoke the following covariance properties of the Radon transform:

1. Rotation covariance. Let $g(x) = g^0(U_{\theta^0}x)$. Then

$$(Rg)(t, \theta) = (Rg^0)(t, \theta - \theta^0).$$

2. Translation covariance. Let $g(x) = g^0(x - x^0)$. Then

$$(Rg)(t, \theta) = (Rg^0)(t - x_1^0 \cos\theta - x_2^0 \sin\theta, \theta).$$

Combining these observations with the formula

$$\alpha_\lambda = [(\Delta^+ \otimes I)Rg, W_\lambda],$$

we see that, with $F(t, \theta) = (\Delta^+ \otimes I)Rg$ and $F^0(t, \theta) = (\Delta^+ \otimes I)Rg^0$, the question of sparsity of the coefficients $(\tilde{\alpha}_\lambda : \lambda \in \Lambda)$ is equivalent to sparsity of wavelet coefficients of

$$F(t, \theta) = F^0(t - x_1^0 \cos(\theta) - x_2^0 \sin(\theta),\ \theta - \theta^0). \tag{6.1}$$

Thus we are interested in the assertion that the smooth change-of-variables,

$$\theta \mapsto \theta - \theta_0; \qquad t \mapsto t - (x_1^0 \cos(\theta) + x_2^0 \sin(\theta)) \tag{6.2}$$

preserves the sparsity of the W_λ wavelet coefficients. To some readers this may seem an innocuous change, but to wavelet experts there would seem to be plenty of reason for this transformation to cause major problems. The W_λ basis is closely connected to a simple tensor product of two wavelet bases, and for such bases, the coefficients are well known to be profoundly affected by very smooth nonlinear changes-of-variables. It turns out *in this case* that the transformation (6.2) has a banal effect on the coefficients because of three interacting factors:

(F1) The transformation of variables is very smooth (C^∞).

(F2) The transformation has a very special structure: it acts on the t variable only by translation according to a function of θ alone, and it acts as a simple shift in the θ variable.

(F3) The basis (W_λ) obeys the constraint $i \geq j$, which is especially compatible with nonlinear transformations with structural feature (F2).

As a result of this combination of factors, we have the following theorem.

THEOREM 6.1. *The coefficients of an F defined by transformation* (6.1) *belong to every ℓ^p for $p > 0$.*

This theorem will complete the proof of Theorem 1.3 of the introduction.

We now give the full argument. Let τ denote the transformation $F = \tau F_0$ defined by (6.1). Let α^0 denote the sequence of ridgelet coefficients of F^0 and α the sequence of ridgelet coefficients of F. Let T denote the matrix with entries $t_{\lambda,\lambda'} = [W_\lambda, \tau W'_\lambda]$. Then, in a pattern of reasoning already familiar from section 5.3,

$$\alpha = T\alpha^0.$$

The desired sparsity of ridgelet coefficients therefore is reduced to the following lemma.

LEMMA 6.2. *T is a bounded mapping from $\ell^p(\Lambda)$ to $\ell^p(\Lambda)$ $\forall$ $p > 0$.*

To prove this, we first observe that in the interesting range $p \in (0, 1]$,

$$\|T\|^p \leq \sup_{\lambda'} \sum_\lambda |t_{\lambda,\lambda'}|^p.$$

Recall now the approach of section 5.3, where properties of a W_λ-wavelet analysis (where antipodal symmetry is imposed on the basis elements) were inferred from a traditional wavelet analysis (without the antipodal symmetry). With $(\tilde{\psi}_\lambda)_{\tilde{\lambda}}$ denoting the orthobasis for $L^2(dtd\theta)$ without antipodal symmetry, suppose we can show the finiteness of

$$\sup_{\tilde{\lambda}'} \sum_{\tilde{\lambda}} |[\tilde{\psi}_{\tilde{\lambda}}, \tau\tilde{\psi}_{\tilde{\lambda}'}]|^p. \tag{6.3}$$

This will then imply finiteness of the norm of T.

Each sum inside the supremum of (6.3) can be interpreted as calculating the (pth power of) the ℓ^p-norm of the coefficients of $\tau\tilde{\psi}_{\tilde{\lambda}'}$. Thus we are interested in the assertion that the smooth deformation of coordinates (6.2) transforms "atoms"—individual basis elements—into "molecules"—sparse sums of basis elements. While such an assertion is not true for arbitrary deformations, it is true for deformations with the special structure considered in (F2) above, when the basis obeys the peculiar constraint $i \geq j$ mentioned in (F3).

To explain this claim, let

$$\Psi_{j,j'}(d) = \int \psi_{j,0}(t)\psi_{j',0}(t-d)dt.$$

Observe that $\Psi_{j,j'} = \Psi_{j-j',0}$; moreover, because Meyer wavelets are being used, $\Psi_{j,j'} = 0$ for $|j - j'| > 1$. Also each $\Psi_{h,0}$, for $h = 0, 1, -1$, is C^∞ and obeys, together with all of its derivatives, rapid decay estimates. Now let $\nu(\theta) = x_1^0 \cos(\theta) + x_2^0 \sin(\theta)$. The (t, θ) integral defining the wavelet coefficients can be reexpressed using $\Psi_{j,j'}$ and ν into a one-dimensional integral of θ alone:

$$\begin{aligned}[\tilde{\psi}_\lambda, \tau\tilde{\psi}_{\lambda'}] &= \frac{1}{4\pi}\int_0^{2\pi} w_{i,\ell}(\theta)w_{i',\ell'}(\theta-\theta_0)\int_{-\infty}^{\infty} \psi_{j,k}(t)\psi_{j',k'}(t-\nu(\theta))dtd\theta \\ &= \frac{1}{4\pi}\int_0^{2\pi} w_{i,\ell}(\theta)w_{i',\ell'}(\theta-\theta_0)\Psi_{j,j'}(t_{j',k'} + \nu(\theta) - t_{j,k})d\theta.\end{aligned}$$

This integral involves three terms: two wavelets and a Ψ-factor. Observe that $i \geq j$, $i' \geq j'$. Also $|j - j'| \leq 1$ in order for $\Psi \neq 0$. The $\Psi_{j,j'}$ factor may therefore be viewed as a smooth function at a scale 2^{-j}. Owing to the $i \geq j$ constraint in forming the basis, the scale of the $\Psi_{j,j'}$ factor *is coarser than the scale of the two wavelet terms $w_{i,\ell}$ and $w_{i,\ell'}$*. Indeed, the $w_{i',\ell'}(\theta + \theta_0)$ factor is either a smooth function at the same scale $2^{-j'}$ ($\approx 2^{-j}$, as $|j - j'| \leq 1$), or an oscillatory one at a finer scale $2^{-i'}$, $i' > j'$.

The $w_{i,\ell}(\theta)$ factor is either a smooth function at the same scale 2^{-j} or an oscillatory function at a finer scale.

We recall the pattern of reasoning of section 5.5, which developed decay estimates for certain inner products, which we now view as integrals involving *two* factors, a wavelet and a deformed wavelet. The pattern was (i) for two nonoscillatory factors at the same scale, use rapid decay of the factors to infer that the integral decays rapidly with increasing separation between the locations of the two factors; (ii) for factors at different scales, observe that because $i \geq j$, the finer scale factor must be oscillatory and the coarser scale factor smooth; using integration by parts, combined with rapid decay estimates on derivatives, show that the integral decays rapidly in the spatial separation of the two factors, as well as rapidly in the scale separation of the two factors.

To apply this pattern of reasoning in the present case, we may group the *three* terms in our integrand into two terms and reason as before. The grouping decision goes by cases, depending on λ and λ'. The cases are Λ_1, where $i = j$ and $i' = j'$ (the coarse-scale case); Λ_2, where $i > i' \geq j$; and Λ_3, where $i' \geq i > j$. We can then show that

$$\sum_{\lambda \in \Lambda_a} |[\tilde{\psi}_\lambda, \tau \tilde{\psi}_{\lambda'}]|^p \leq C$$

with constant C independent of λ' and of $a \in \{1, 2, 3\}$.

In the first case, we argue from rapid decay as in case $S_{0,0}$ of section 5.5. In Λ_2 and Λ_3, we argue that the product of the Ψ term with the coarser-scale wavelet (either $w_{i,\ell}(\theta)$ or $w_{i',\ell'}(\theta + \theta_0)$ as the case may be) yields a factor which obeys the same smoothness and localization bounds that the coarser-scale wavelet obeys. We then use the integration-by-parts argument of cases $S_{0,1}$ and $S_{1,0}$ in section 5.5. This completes the proof of Lemma 6.2.

7. Discussion.

7.1. Alternate ridgelet orthobasis. Another natural construction of orthobasis can be made using the ideas of sections 2–4. Define the index set

$$\begin{aligned} \Lambda' = & \{(j,k;i,l,0) : i = i_0; j,k \in \mathbb{Z};\ \ell = 0, \dots, 2^{i-1} - 1\} \\ & \cup \quad \{(j,k;i,l,1) : i \geq i_0; j,k \in \mathbb{Z};\ \ell = 0, \dots, 2^{i-1} - 1\}. \end{aligned} \tag{7.1}$$

In comparison with the set Λ, notice that we may have either $j > i$ or $j \leq i$, and that $\varepsilon = 0$ is only compatible with $i = i_0$. It is easy to see that $(W_\lambda : \lambda \in \Lambda')$ is a complete orthonormal system for $\mathcal{R}$, and so the isometry $\rho_\lambda = \mathcal{J}(W_\lambda)$ makes $(\rho_\lambda : \lambda \in \Lambda')$ a complete orthonormal system for $L^2(\mathbb{R}^2)$.

This alternate system of orthonormal ridgelets has an attractive "angular multiresolution" interpretation where angular behavior over coarser scales than 2^{-j} is represented in the transform. In the technical report version of this article [8], this basis was studied carefully. For reasons of space, we have omitted that analysis here.

7.2. Alternate *g*. Obviously the approach we have developed is not limited to the Gaussian case studied here. A natural class of examples to study is the form $g^0(x_1, x_2) = g_1(x_1) \cdot g_2(x_2)$, where $g_1 \in \mathcal{S}(\mathbb{R})$ and $g_2(x_2)$ is smooth away from a singularity at $x_2 = 0$, and on each half line is of rapid decay along with its derivatives. The simple case $g_2(x_2) = e^{-x_2} 1_{\{x_2 > 0\}}$ is rather easy to study and yields the same qualitative conclusions as in the Gaussian case.

The Fourier-domain estimation technique that we have developed here requires a considerable amount of effort to carry out. A lot of the effort is based on the fact that we are studying objects of unlimited smoothness away from the singularity, and we want to show that the rate of decay of the ridgelet coefficients is unlimited. For general results on functions of limited smoothness, assuming only qualitative properties of g, such as Hölder smoothness of order m, the task is in principle less challenging, because one hopes only to establish a limited rate of decay of the ridgelet coefficients. It would be useful to have an easier technique for coefficient estimation.

The report [8] developed an approach to estimation of ridgelet coefficients using Radon-domain estimates, rather than Fourier-domain estimates, and wavelets of compact support [4], rather than Meyer wavelets. That approach could be useful in dealing with objects with limited smoothness, and might well be easier to apply in certain situations.

7.3. Curvilinear singularity. We stress that orthonormal ridgelet analysis of an object with singularity along a curve does not, in general, yield sparse coefficients.

In effect, the Radon transform of such an object has a singularity along a curve, and not just at a point. Consider, for example, the object $g' = e^{-x_1^2-x_2^2} \cdot 1_{\{x_2 > x_1^2\}}$. Define $\Phi(a,b) = \int_a^b e^{-u^2} du$. Using the geometric viewpoint of Figure 8.1 in the appendix, with coordinates t and u rotated by angle θ from the coordinates x_1 and x_2, one can write the Radon transform Rg' as

$$(Rg')(t,\theta) = e^{-t^2} \Phi(u_-(\theta, t - t_0(\theta)), u_+(\theta, t - t_0(\theta))),$$

where $t_0(\theta)$ is the t-coordinate for which $\mathcal{L}_{(\theta,t)}$ is tangent to the parabola $x_2 = x_1^2$, and, with orientation chosen so that for $t > t_0$, $\mathcal{L}_{(\theta,t)}$ intersects the parabola, the functions $u_\pm$ are smooth functions of θ and $t - t_0$ giving the u-coordinates of the two points at which the line $\mathcal{L}_{(\theta,t)}$ intersects the parabola. The function $t_0(\theta)$ is a smooth function of θ, and for an appropriate smooth function $c(\theta)$, we have the asymptotic

$$u_+(\theta, \delta t) - u_-(\theta, \delta t) \sim c(\theta)|\delta t|^{1/2} \qquad \text{as } |\delta t| \to 0.$$

Letting $u_0(\theta)$ be the common limit of $u_\pm(\theta, \delta t)$ as $t \to t_0(\theta)$, we therefore have

$$\Phi(u_-(\theta, t - t_0(\theta)), u_+(\theta, t - t_0(\theta))) \sim e^{-u_0^2} \cdot c(\theta)|\delta t|^{1/2} \qquad \text{as } |\delta t| \to 0.$$

Hence Rg' has a singularity of order $1/2$ in the t-direction at points $(\theta, t_0(\theta))$ lying along a smooth curve.

As orthonormal ridgelet analysis amounts to a kind of wavelet analysis in the Radon domain, and as wavelet analysis of singularities along curves does not yield sparse coefficients, so the ridgelet coefficients of such an object are not sparse. The ridgelet coefficients of such a curved object decay, in general, no faster than the wavelet coefficients of the same object. However, with an appropriate multiscale localization of the ridgelet basis, a significant improvement over wavelet analysis can be obtained.

7.4. Higher dimensions. Our construction of orthonormal ridgelets relies on two facts: first, the existence of orthonormal wavelets on the circle, and second, the existence of an isometry between antipodally symmetric functions in Radon space and functions in Real space. To obtain the analogous construction straightforwardly in dimensions $d > 2$, we would need orthonormal wavelets on the sphere S^{d-1} and an isometry for higher dimensions. The isometry exists in every dimension $d \geq 2$ [12].

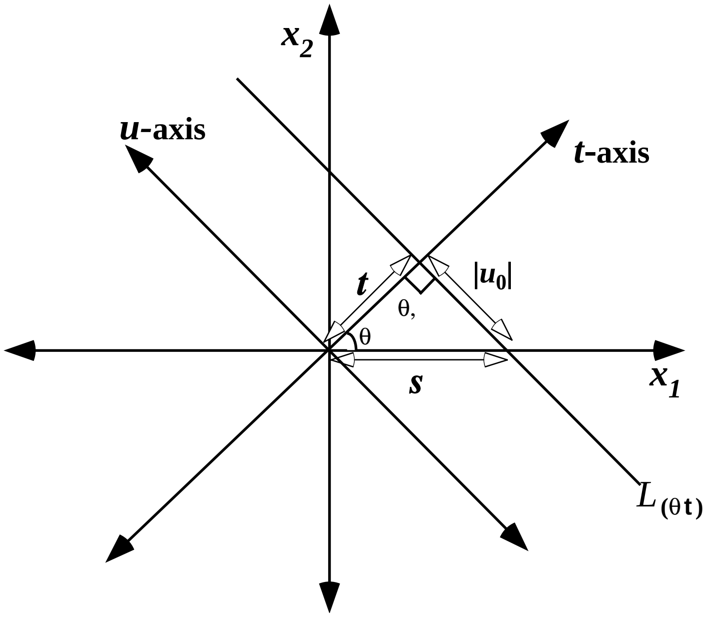


FIG. 8.1. *Geometry for calculating the Radon transform of g.*

Unfortunately, orthonormal spherical wavelets are not known for any dimension $d > 2$. The next best thing to an orthonormal system is a tight frame, which obeys a Parseval relation. The article [10] constructs tight frames of wavelet-like elements on spheres of all dimensions and so obtains tight frames of ridgelets in all dimensions $d > 2$. It also shows how to construct k-plane ridgelets (tight frame expansions substituting for functions depending on k-variables) for $k = 1, \ldots, d-1$ in every dimension $d > 2$. It is an interesting question whether results paralleling Theorem 1.3 can be had in higher dimensions d and for various codimensions k. [10] suggests that the answer is yes whenever $k = d - 1$.

8. Appendix.

8.1. Radon transform of *g*. The diagram in Figure 8.1 shows how to derive the Radon transform of g.

The Radon transform of g at (t, θ) may be viewed the integral of $e^{-x_1^2 - x_2^2}$ along that part of the line $\mathcal{L}_{(\theta,t)}$ lying inside the upper halfplane. For θ fixed, introduce orthogonal coordinates (t, u) with t the same as the t variable in the Radon transform. Thus the line of integration in the Radon transform is expressible as $\mathcal{L}_{(\theta,t)} = \{(t,u) : u \in \mathbb{R}\}$ in the new coordinates. Let u_0 denote the least value of u for which (t, u) is in both $\mathcal{L}_{(\theta,t)}$ and the upper halfplane $x_2 \geq 0$. Also, note that by orthogonality of the coordinates, $e^{-x_1^2 - x_2^2} = e^{-t^2 - u^2}$. Hence

$$(Rg)(t,\theta) = \int_{u_0}^{\infty} e^{-t^2-u^2} du = e^{-t^2} \cdot \int_{u_0}^{\infty} e^{-u^2} du = e^{-t^2} \cdot \bar{\Phi}(u_0),$$

say, where here and below $\bar{\Phi}(v) \equiv \int_v^\infty e^{-u^2} du$. Now from the geometry of the figure we can see that there is a right triangle with hypotenuse s, say, so that $|t|$ and $|u_0|$

are the lengths of the other two sides, and so

$$|t| = s\cos(\theta); \qquad |u_0| = s\sin(\theta).$$

In the range $0 \le \theta < \pi/2$, $u_0 = -\sin(\theta)/\cos(\theta)\cdot t$. Similar diagrams for other ranges of θ show that $u_0 = -|\sin(\theta)/\cos(\theta)|\cdot t$ is the correct general formula. We therefore have (5.2). $\square$

8.2. Fourier transform of $\boldsymbol{S}$. The key formula (5.5) is essentially an immediate application of the following lemma, which will be proved in a moment.

LEMMA 8.1. *Let $\dot{H}(u,v) = \bar{\Phi}(u/|v|)$. This bounded function, viewed as tempered distribution, has Fourier transform*

$$\widehat{\dot{H}} = 2\pi^{5/2}\delta_0 - 2\pi\cdot\sqrt{-1}\cdot\mathrm{sgn}(\omega_1)|\omega_1|^{-2}e^{-\omega_2^2/\omega_1^2}, \tag{8.1}$$

where δ_0 denotes the Dirac mass at $(\omega_1,\omega_2) = (0,0)$.

To use this to get (5.5), write the formula in (8.1) as $\widehat{\dot{H}} = \gamma_1\delta_0 + \widehat{\dot{H}}_1(\omega_1,\omega_2)$, with constant γ_1. Comparing the formulas in the two lemmas, we have that $\hat{S}(\omega_1,\omega_2) = \widehat{\dot{H}}_1(\omega_1,\omega_2)|\omega_1|^{1/2}$. In other words, we form $\hat{S}$ by multiplying the proper function $\widehat{\dot{H}}_1$ by $|\omega_1|^{1/2}$, while completely ignoring the singular term supported at 0. This works because

$$\begin{aligned}
\langle S, f\rangle &\equiv \langle \dot{H}, (\Delta^+\otimes I)f\rangle\\
&= \frac{1}{4\pi^2}\langle \widehat{\dot{H}}, ((\Delta^+\otimes I)f)\hat{}\,\rangle\\
&= \frac{1}{4\pi^2}\Big(\gamma_1((\Delta^+\otimes I)f)\hat{}\,(0,0)\\
&\qquad + \iint \widehat{\dot{H}}_1(\omega_1,\omega_2)|\omega_1|^{1/2}\hat{f}(\omega_1,\omega_2)d\omega_1 d\omega_2\Big)\\
&= \frac{1}{4\pi^2}\iint \widehat{\dot{H}}_1(\omega_1,\omega_2)|\omega_1|^{1/2}\hat{f}(\omega_1,\omega_2)d\omega_1 d\omega_2\\
&\equiv \frac{1}{4\pi^2}\langle \hat{S}, \hat{f}\rangle;
\end{aligned}$$

the singular term in $\widehat{\dot{H}}$ supported at the origin never enters because of the vanishing of $\hat{f}$ and $|\omega_1|^{1/2}$ there.

It is enough to establish the formula (8.1) for tensor products $f\otimes g$, i.e.,

$$\begin{aligned}
\langle \dot{H}, f\otimes g\rangle &= \frac{1}{4\pi^2}\langle 2\pi^{5/2}\delta_0, (f\otimes g)\hat{}\,\rangle\\
&\quad + \frac{1}{4\pi^2}\left\langle \frac{2\pi}{\sqrt{-1}}\frac{\mathrm{sgn}(\omega_1)}{|\omega_1|^2}e^{-\omega_2^2/\omega_1^2}, (f\otimes g)\hat{}\,\right\rangle.
\end{aligned} \tag{8.2}$$

We begin with some one-dimensional Fourier analysis. $\bar{\Phi}$ is a bounded function, and as a tempered distribution has, for $f\in\mathcal{S}$,

$$\int_{-\infty}^{\infty}\bar{\Phi}(t)f(t)dt = \frac{\sqrt{\pi}}{2}\int_{-\infty}^{\infty}f(t)dt - \frac{\sqrt{-1}}{2\pi}\int_{-\infty}^{\infty}\frac{\hat{\phi}(\omega)}{\omega}\hat{f}(\omega)d\omega, \tag{8.3}$$

where $\phi(t) = e^{-t^2}$ and where the last integral should be interpreted as

$$\int_0^\infty \hat{\phi}(\omega)(\hat{f}(\omega) - \hat{f}(-\omega))/\omega d\omega,$$

which is absolutely convergent for $f \in \mathcal{S}$. The justification of (8.3) can be obtained either by viewing $\bar{\Phi}(-t)$ as the convolution of a Heaviside function with ϕ, and using the formula for the Fourier transform of a Heaviside [11, p. 172] or directly as follows.

$\bar{\Phi}(t)$ is of the form $\frac{\sqrt{\pi}}{2} + \nu(t)$, where $\nu(t)$ is an odd bounded C^∞ function. Hence, splitting $f = f_e + f_o$ into its even and odd parts,

$$\int_{-\infty}^\infty \bar{\Phi}(t) f(t) dt = \frac{\sqrt{\pi}}{2} \int_{-\infty}^\infty f_e(u) du + \int_{-\infty}^\infty \nu(t) f_o(t) dt.$$

From $f_o \in \mathcal{S}$ and integration by parts,

$$\int_{-\infty}^\infty \nu(t) f_o(t) dt = -\int_{-\infty}^\infty \nu'(t) F_o(t) dt = \int_{-\infty}^\infty \phi(t) F_o(t) dt,$$

where $-\nu' = -(\bar{\Phi})' = \phi$, and F_o is the primitive of f_0. For an odd function $f_o \in \mathcal{S}$, the primitive F_o, viewed as a tempered distribution, has a Fourier transform which is represented by integration against a proper function, and obeys the formula

$$\hat{F}_o(\omega) = (i\omega)^{-1} \hat{f}_o(\omega).$$

Hence

$$\int_{-\infty}^\infty \phi(t) F_o(t) dt = \frac{1}{2\pi} \int \hat{\phi}(\omega) \cdot (i\omega)^{-1} \hat{f}_o(\omega) d\omega.$$

Note, however, that as f is real and f_e is even, $\hat{f}_e$ is even. This integral therefore is insensitive to the difference between $\hat{f}$ and $\hat{f}_o$:

$$\frac{1}{2\pi} \int \hat{\phi}(\omega) \cdot (i\omega)^{-1} \hat{f}(\omega) d\omega = \frac{1}{2\pi} \int \hat{\phi}(\omega) \cdot (i\omega)^{-1} \hat{f}_o(\omega) d\omega;$$

(8.3) follows.

Note now that

$$\int_{-\infty}^\infty \bar{\Phi}(u/|v|) f(u) du = \frac{\sqrt{\pi}}{2} \int_{-\infty}^\infty f(u) du - \frac{\sqrt{-1}}{2\pi} \int_{-\infty}^\infty \frac{\hat{\phi}(v\omega_1)}{\omega_1} \hat{f}(\omega_1) d\omega_1,$$

and so

$$\begin{aligned} \int_{-\infty}^\infty g(v) \int_{-\infty}^\infty \bar{\Phi}(u/|v|) f(u) du dv &= \frac{\sqrt{\pi}}{2} \Big(\int_{-\infty}^\infty f(u) du \Big) \Big(\int_{-\infty}^\infty g(v) dv \Big) \\ &\quad - \frac{\sqrt{-1}}{2\pi} \int_{-\infty}^\infty g(v) \int_0^\infty \hat{\phi}(v\omega_1) h(\omega_1) d\omega_1 dv \\ &= I + II, \end{aligned}$$

say, where $h(\omega_1) \equiv (\hat{f}(\omega_1) - \hat{f}(-\omega_1))/\omega_1$. Now as $\hat{\phi}$ is bounded and g and h are absolutely integrable on their domains, Fubini applies; hence

$$II = -\frac{\sqrt{-1}}{2\pi} \int_0^\infty h(\omega_1) \int_{-\infty}^\infty g(v) \hat{\phi}(v\omega_1) dv d\omega_1.$$

Parseval gives, for $\omega_1 \neq 0$,

$$\int_{-\infty}^{\infty} g(v)\hat{\phi}(v\omega_1)dv = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{g}(\omega_2)\tilde{\phi}_{\omega_1}(\omega_2)d\omega_2,$$

where

$$\begin{aligned}\tilde{\phi}_{\omega_1}(\omega_2) &\equiv \int_{-\infty}^{\infty} \hat{\phi}(v\omega_1)e^{-iv\omega_2}dv \\ &= \frac{2\pi}{|\omega_1|}e^{-\omega_2^2/\omega_1^2}.\end{aligned}$$

Now

$$\frac{2\pi}{|\omega_1|}e^{-\omega_2^2/\omega_1^2}$$

is in $L^1((0,\infty)\times\mathbb{R})$; indeed, it is homogeneous of degree -1 and is integrable along the four line segments on the boundary of the unit square. It follows that we may unambiguously write

$$II = -\frac{\sqrt{-1}}{(2\pi)^2}\int_{-\infty}^{\infty}\int_0^{\infty} \hat{g}(\omega_2)h(\omega_1)\frac{2\pi}{|\omega_1|}e^{-\omega_2^2/\omega_1^2}d\omega_1 d\omega_2.$$

Unwrapping the formula for h in terms of $\hat{f}$, this becomes the second term of the desired formula (8.2). As for the first term in that formula, this follows from

$$I = \frac{\sqrt{\pi}}{2}\left(\int_{-\infty}^{\infty} f(u)du\right)\left(\int_{-\infty}^{\infty} g(v)dv\right) = \frac{\sqrt{\pi}}{2}\hat{f}(0)\hat{g}(0) = \frac{\sqrt{\pi}}{2}\langle\delta_0, \hat{f}\otimes\hat{g}\rangle. \qquad \square$$

8.3. Proof of Lemma 5.4. Define

$$\beta_t(\omega_1,\omega_2) = t\ e^{-t^4 E(\omega_1,\omega_2)},$$

where the exponent function E obeys

[E1] $E(\omega_1,\omega_2) \in C^m[\Omega]$, $m = 1,2,3,\ldots$; and

[E2] $|E(\omega_1,\omega_2)| \geq C > 0$ on Ω.

The particular case $E(\omega_1,\omega_2) = \omega_2^2/\omega_1^2$ gives $B_h = \beta_t$ with $t = 2^{h/2}$. This does have the required properties [E1]–[E2] because on Ω, $\frac{2}{3}\pi \leq |\omega_i| \leq \frac{8}{3}\pi$, and so the ratio stays well away from zero and infinity.

We will show that

$$\|\beta_t\|_{C^m[\Omega]} \leq 2^{-\lambda t}\cdot \text{const}, \qquad t > t(\lambda), \tag{8.4}$$

giving (5.12). Indeed

$$\left(\frac{\partial}{\partial\omega_1}\right)^{m_1}\left(\frac{\partial}{\partial\omega_2}\right)^{m_2}\beta_t = e^{-t^4E}\left\{\sum_{d=0}^{D_0(m_1,m_2)} P_d(t)Q_d(E, E^{(1,0)},\ldots)\right\},$$

where each P_d is a polynomial of degree $\leq D_1(m)$, each Q_d is a multivariate polynomial of degree $\leq D_2(m)$, and where the degrees in question satisfy $D_i(m) \leq C_1 + C_2(m_1 + m_2)$, $i = 0,1,2$. Here the $E^{(m_1,m_2)}$ are mixed partial derivatives of E. Hence for an $m' \leq C_1 + C_2 m$ we have

$$\|\beta_t\|_{C^m[\Omega]} \leq e^{-t^4E}\tilde{P}(t)\tilde{Q}(\|E\|_{C^{m'}[\Omega]}),$$

where $\tilde{P}$ and $\tilde{Q}$ are univariate polynomials of degree $\leq D_3(m) \leq C_1' + C_2'(m)$.

Now under [E2], $|E| > C$, and so for $\lambda > 0$, there is $t(\lambda)$ so $e^{-t^4C} < 2^{-\lambda t}$ $\forall$ $t > t(\lambda)$, giving (8.4), valid $\forall$ $\lambda > 0$. $\square$

8.4. Proof of Lemma 5.10. Throughout this section, let

$$\partial^{(m,n)} \equiv \left(\frac{\partial}{\partial \xi_1}\right)^m \left(\frac{\partial}{\partial \xi_2}\right)^n .$$

8.4.1. Auxiliary lemmas.

LEMMA 8.2. *Let* $\beta(\omega_1, \omega_2) = \mathrm{sgn}(\omega_1)|\omega_1|^{-2}\exp\{-\frac{\omega_2^2}{\omega_1^2}\}$. *Then* $\beta \in C^\infty(\mathbb{R} \times ([-8\pi/3, -2\pi/3] \cup [2\pi/3, 8\pi/3]))$. *Also*

$$\left(\frac{\partial}{\partial \omega_1}\right)^m \left(\frac{\partial}{\partial \omega_2}\right)^n \beta(\omega_1, \omega_2) = \mathrm{sgn}(\omega_1)^{m+n+1}|\omega_1|^{-2-m-n}F_{m,n}(\omega_2/\omega_1), \tag{8.5}$$

where $F_{m,n} \in \mathcal{S}(\mathbb{R})$.

Proof. Let $F \in \mathcal{S}(\mathbb{R})$. On $\omega_1 \neq 0$,

$$\frac{\partial}{\partial \omega_1}\left[|\omega_1|^{-2-\ell}F(\omega_2/\omega_1)\right] = \mathrm{sgn}(\omega_1)|\omega_1|^{-2-\ell-1}\tilde{F}(\omega_2/\omega_1),$$

where $\tilde{F}(t) = (-2-\ell)F(t) - F'(t)\cdot t$ satisfies $\tilde{F} \in \mathcal{S}(\mathbb{R})$. Similarly, on $\omega_1 \neq 0$,

$$\frac{\partial}{\partial \omega_2}\left[|\omega_1|^{-2-\ell}F(\omega_2/\omega_1)\right] = \mathrm{sgn}(\omega_1)|\omega_1|^{-2-\ell-1}\tilde{F}(\omega_2/\omega_1),$$

where now $\tilde{F}(t) = F'(t)$ and again $\tilde{F} \in \mathcal{S}(\mathbb{R})$. Repetitively applying this pair of observations gives (8.5). □

LEMMA 8.3.

$$|\partial^{(m,n)}\nu_1\beta| \leq C_{m,n}\cdot(1+|\omega_2|)^{-m-n-2}, \quad \omega \in \mathbb{R}^2. \tag{8.6}$$

Proof. First, we remark that when $\omega_1, \omega_2 \neq 0$,

$$\beta^{(m,n)}(\omega_1, \omega_2) = |\omega_1|^{-m-n-2}\cdot \mathrm{sgn}(\omega_1)^{m+n+1}\cdot F_{m,n}(\omega_2/\omega_1),$$

where $F_{m,n} \in \mathcal{S}(\mathbb{R})$; see Lemma 8.2 above. Second, putting now $\nu = \nu_1$, there are constants $C_{\ell,m}$ so that

$$\begin{aligned}\partial^{(m,n)}\nu\beta &= \partial^{(m,0)}(\nu\cdot\partial^{(0,n)}\beta)\\ &= \sum_{\ell=0}^{m} C_{\ell,m}\nu^{(\ell,0)}\beta^{(m-\ell,n)},\end{aligned}$$

where $\nu^{(\ell,0)} \equiv \partial^{(\ell,0)}\nu$. Now for $\ell > 0$, $\nu^{(\ell,0)}$ is supported in $\{\omega : |\omega_1| \in [1,2]\}$. Hence, there is $\tilde{\nu}_\ell \in \mathcal{S}(\mathbb{R}^2)$ so that

$$\nu^{(\ell,0)}(\omega) = \tilde{\nu}_\ell(\omega)\cdot \mathrm{sgn}(\omega_1)^\ell|\omega_1|^{-\ell}.$$

We have

$$|\partial^{(m,n)}\nu\beta| \leq \sum_{\ell=0}^{m}|\tilde{\nu}_\ell|\cdot|\omega_1|^{-m-n-2}|F_{n,m,\ell}(\omega_2/\omega_1)|$$

for $F_{n,m,\ell} \in \mathcal{S}(\mathbb{R})$. Now as each $F_{n,m,\ell} \in \mathcal{S}(\mathbb{R})$, we have

$$F_{n,m,l}(t) \leq C\cdot|t|^{-m-n-2} \qquad \forall|t| > 1.$$

On the set where $|\omega_2| > |\omega_1|$, this gives the inequality

$$|\omega_1|^{-m-n-2}|F_{n,m,\ell}(\omega_2/\omega_1)| \leq C \cdot |\omega_2|^{-m-n-2}$$

and on the set where $|\omega_2| < |\omega_1|$, $F_{n,m,\ell} \in \mathcal{S}(\mathbb{R})$ gives the inequality

$$|\omega_1|^{-m-n-2}|F_{n,m,\ell}(\omega_2/\omega_1)| \leq C \cdot |\omega_1|^{-m-n-2} \leq C \cdot |\omega_2|^{-m-n-2}.$$

Now on $|\omega_2| \geq 1$, $|\omega_2|^{-m-n-2} \leq C(1+|\omega_2|)^{-m-n-2}$. Equation (8.6) follows from this and supp $\nu_\ell \subset ([-1,1]^2)^c$. □

LEMMA 8.4. *For $\mu > \alpha + 1$, $\mu > 2$, $\alpha > 0$*

$$\int (1+|\omega|)^{-\mu}(1+|\xi-\omega|)^{-\alpha} d\omega \leq C_{\mu,\alpha}(1+|\xi|/2)^{-\alpha}. \tag{8.7}$$

Proof. Suppose without loss of generality that $\xi > 1$.

$$\begin{aligned}\int_{-\infty}^{\xi/2} (1+|\omega|)^{-\mu}(1+|\xi-\omega|)^{-\alpha} d\omega &\leq (1+|\xi|/2)^{-\alpha} \cdot \int_{-\infty}^{\xi/2} (1+|\omega|)^{-\mu} d\omega \\ &\leq C_\mu \cdot (1+|\xi|/2)^{-\alpha}.\end{aligned}$$

$$\begin{aligned}\int_{\xi/2}^{\infty} (1+|\omega|)^{-\mu}(1+|\xi-\omega|)^{-\alpha} d\omega &\leq \int_{\xi/2}^{\infty} (1+|\omega|)^{-\mu} d\omega \\ &= (\xi/2+1)^{1-\mu}/(\mu-1) \\ &\leq C_{\mu,\alpha}(1+|\xi|/2)^{-\alpha}, \qquad \xi > 1, \mu > \alpha+1. \qquad □\end{aligned}$$

8.4.2. Proof of (5.19). Now let $\Omega_j^h = \{(2^j \xi_1, 2^{j+h}\xi_2) : \xi \in \Omega\}$. Then from (5.17) and (5.18) we get the scaling relation

$$\|\partial^{(m,n)} B_{j,h}\|_{L^\infty[\Omega]} = (2^h)^{(n+1/2)} 2^{(m+n+3/2)j} \|\beta_0^{(m,n)}\|_{L^\infty[\Omega_j^h]}. \tag{8.8}$$

Our proof of (5.19) shows that

$$\|\beta_0^{(m,n)}\|_{L^\infty[\Omega_j^h]} \leq C_{m,n} 2^{-h(n+2)} 2^{-j(n+m+2)}; \tag{8.9}$$

it follows from this and (8.8) that

$$\|\partial^{(m,n)} B_{j,h}\|_{C^m[\Omega]} \leq C_{m,n} \cdot 2^{-h(3/2)} 2^{-j(1/2)}, \quad m = 1,2,3,\ldots,$$

from which the summability (5.19) follows immediately.

To analyze $\beta_0^{(m,n)}$, we partition the integration into two pieces via smooth windows. Let $\nu_r(\omega_1,\omega_2)$, for $r = 0,1$, denote smooth bivariate separable windows obeying $0 \leq \nu_r \leq 1$, $\nu_0(\omega) + \nu_1(\omega) \equiv 1$, $\nu_0 \equiv 1$ on $[-1,1]^2$, and $\nu_1 \equiv 1$ on $([-2,2]^2)^c$, and ν_0 and ν_1 both even in both coordinates. Now with $(\nu_r\beta)(\omega_1,\omega_2) \equiv \nu_r(\omega_1,\omega_2)\beta(\omega_1,\omega_2)$ define

$$T_0^r(\xi_1,\xi_2) = \iint V_0(\omega_1,\omega_2)(\nu_r\beta)(\xi_1-\omega_1,\xi_2-\omega_2) d\omega_1 d\omega_2,$$

so that

$$\beta_0 = T_0^0 + T_0^1,$$

T_0^1 containing contributions to β_0 arising far from $\omega = 0$, and T_0^0 containing contributions arising $\omega = 0$. Again T_0^0 involves a nonabsolutely convergent integrand, and part of the work will be to make sense of it.

About these terms we make the following claims. First, that for $m, n = 0, 1, 2, \ldots$, and $\mu > 0$,

$$|\partial^{(m,n)} T_0^0(\xi_1, \xi_2)| \leq C_{m,n,\mu} \cdot (1 + (|\xi_2| - 2)_+)^{-\mu}, \tag{8.10}$$

where $C_{m,n,\mu}$ does not depend on ξ. Second, that

$$|\partial^{(m,n)} T_0^1(\xi_1, \xi_2)| \leq C'_{m,n}\ (1 + |\xi_2|/2)^{-m-n-2}, \tag{8.11}$$

where $C'_{m,n}$ does not depend on ξ.

Combining these claims, if $\xi \in \Omega_j^h$, then for $r = 0, 1$,

$$|\partial^{(m,n)} T_0^r(\xi_1, \xi_2)| \leq C_{m,n}(1 + (2^{j+h}\frac{\pi}{3} - 2)_+)^{-m-n-2},$$

from which (8.9) follows and so the conclusion (5.19) of Lemma 5.10 is established.

We now justify the claims. Separability of the kernel ν_0 allows us to write $\nu_0(\xi_1, \xi_2) = \nu_{0,1}(\xi_1) \cdot \nu_{0,2}(\xi_2)$, and also $V_0(\xi_1, \xi_2) = V_{0,1}(\xi_1) \cdot V_{0,2}(\xi_2)$. Putting $v_1(\xi) = (\frac{\partial}{\partial \xi})^m V_{0,1}(\xi)$ and similarly for v_2, then

$$\begin{aligned}
&\iint V^{(m,n)}(\xi_1 - \omega_1, \xi_2 - \omega_2)\nu_0 \beta d\omega_1 d\omega_2 \\
&= \int_{-\infty}^{\infty} v_2(\xi_2 - \omega_2)\nu_{0,2}(\omega_2) \int_{-\infty}^{\infty} v_1(\xi_1 - \omega_1)\nu_{0,1}(\omega_1)|\omega_1|^{-2}\mathrm{sgn}(\omega_1)e^{-\omega_2^2/\omega_1^2} d\omega_1 d\omega_2 \\
&= \int_{-\infty}^{\infty} v_2(\xi_2 - \omega_2)\nu_{0,2}(\omega_2) \int_0^2 (v_1(\xi_1 - \omega_1) - v_1(\xi_1 + \omega_2))|\omega_1|^{-1} \\
&\qquad \cdot \nu_{0,1}(\omega_1)|\omega_1|^{-1} e^{-\omega_2^2/\omega_1^2} d\omega_1 d\omega_2.
\end{aligned} \tag{8.12}$$

This identity allows us to interpret $\beta_0^{(m,n)}$ as an unconditionally convergent integral. Indeed, for each $\mu > 0$, we have the following inequality valid $\forall\ \xi_1 \in \mathbb{R}$, $\omega_1 \neq 0$:

$$|v_1(\xi_1 - \omega_1) - v_1(\xi_1 + \omega_2)|/|\omega_1| \leq C_\mu \left((1 + |\xi_1 - \omega_1|)^{-\mu} + (1 + |\xi_1 + \omega_1|)^{-\mu}\right).$$

Hence, we can "cancel one factor of $|\omega_1|^{-1}$," yielding pointwise domination of the integrand defining $\beta_0^{(m,n)}$ in the final member in (8.12) by an integrable function, leading to the interpretation of the initial member as bounded according to

$$\begin{aligned}
&\left|\iint V^{(m,n)}(\xi_1 - \omega_1, \xi_2 - \omega_2)\nu_0 \beta d\omega_1 d\omega_2\right| \\
&\leq \int_{-\infty}^{\infty} |v_2(\xi_2 - \omega_2)\nu_{0,2}(\omega_2)| \int_{-\infty}^{\infty} C(1 + |\xi_1 - \omega_1|)^{-\mu}\nu_{0,1}(\omega_1)|\omega_1|^{-1}e^{-\omega_2^2/\omega_1^2} d\omega_1 d\omega_2.
\end{aligned}$$

Now because the window ν_0 is supported in $[-2, 2]^2$, this last term can be bounded by

$$\begin{aligned}
&C \cdot \int_{-2}^{2}\int_{-2}^{2} |\omega_1|^{-1}e^{-\omega_2^2/\omega_1^2} d\omega_1 d\omega_2 \times \sup_{\omega \in [-2,2]^2} |v_2(\xi_2 - \omega_2)|(1 + |\xi_1 - \omega_1|)^{-\mu} \\
&\quad \leq C \cdot (1 + (|\xi_1| - 2)_+)^{-\mu} \cdot (1 + (|\xi_2| - 2)_+)^{-\mu},
\end{aligned}$$

yielding (8.10).

Now consider (8.11). Inside the support of ν_1, we may differentiate $\nu_1\beta$ as often as we like. Applying Lemma 8.3,

$$|\partial^{(m,n)}\nu_1\beta(\omega_1,\omega_2)| \leq C_{m,n,\mu}\cdot(1+|\omega_2|)^{-m-n-2}$$

so that from the integration by parts

$$\iint V^{(m,n)}(\xi_1-\omega_1,\xi_2-\omega_2)\nu_1\beta d\omega_1 d\omega_2 = (-1)^{m+n}\iint V(\xi_1-\omega_1,\xi_2-\omega_2)(\nu_1\beta)^{(m,n)}d\omega_1 d\omega_2$$

and an argument paralleling the proof of Lemma 8.4, we have

$$\begin{aligned}&\left|\iint V^{(m,n)}(\xi_1-\omega_1,\xi_2-\omega_2)\nu_1\beta d\omega_1 d\omega_2\right|\\ &\leq \iint(1+|\xi_1-\omega_1|)^{-\mu}\cdot(1+|\xi_2-\omega_2|)^{-\mu}\cdot(1+|\omega_2|)^{-m-n-2}d\omega_1 d\omega_2\\ &\leq C\cdot(1+|\xi_2|/2)^{-m-n-2}.\end{aligned}$$

Acknowledgment. The author would like to thank Emmanuel Candès for many stimulating discussions.

REFERENCES

[1] E. J. Candès, *Harmonic analysis of neural networks,* Appl. Comput. Harmon. Anal., 6 (1999), pp. 197–218.

[2] E. J. Candès, *Ridgelets: Theory and Applications,* Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, CA, 1998.

[3] E. J. Candès and D. L. Donoho, *Ridgelets: The key to high-dimensional intermittency?,* Philos. Trans. Roy. Soc. London Ser. A, to appear.

[4] I. C. Daubechies, *Orthonormal bases of compactly supported wavelets,* Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[5] S. R. Deans, *The Radon Transform and Some of Its Applications,* Reprinted Ed., Krieger, Malabar, FL, 1991.

[6] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?* J. Roy. Statist. Ser. B, 57 (1995), pp. 301–369.

[7] D. L. Donoho, M. Vetterli, I. Daubechies, and R. A. DeVore, *Data compression and harmonic analysis,* IEEE Trans. Inform. Theory, 44 (1998), pp. 2435–2476.

[8] D. L. Donoho, *Orthonormal Ridgelets and Linear Singularities,* Technical Report, Department of Statistics, Stanford University, Stanford, CA, 1998; also available online from http://www-stat.stanford.edu/˜donoho/Reports/1998/ridge-lin-sing.ps.

[9] D. L. Donoho, *Ridge Functions and Orthonormal Ridgelets,* Technical Report, Department of Statistics, Stanford University, Stanford, CA, 1998; also available online from http://www-stat.stanford.edu/˜donoho/Reports/1998/Ridge-Ridgelet.ps.

[10] D. L. Donoho, *Tight Frames of k-plane ridgelets, and the problem of representing functions smooth away from d-dimensional singularities in $\mathbb{R}^n$,* Proc. Nat. Acad. Sci. USA, 96 (1999), pp. 1828–1833.

[11] I. M. Gel'fand and G. E. Shilov, *Generalized Functions: Properties and Operations,* Academic Press, New York, London, 1964.

[12] S. Helgason, *Groups and Geometric Analysis*, Academic Press, New York, London, 1986.

[13] B. F. Logan and L. A. Shepp, *Optimal reconstruction of a function from its projections,* Duke Math. J., 42 (1975), pp. 645–659.

[14] P. G. Lemarié and Y. Meyer, *Ondelettes et bases Hilbertiennes*, Rev. Mat. Iberoamericana, 2 (1986), pp. 1–18.

[15] R. Strichartz, *A Guide to Distribution Theory and Fourier Transforms,* CRC Press, Boca Raton, FL, 1993.

SIAM J. MATH. ANAL.
Vol. 31, No. 5, pp. 1100–1120

RESOLUTION IN DYNAMIC EMISSION TOMOGRAPHY*

JEAN MAEGHT† AND DOMINIKUS NOLL†

Abstract. Based on a two-dimensional (2-D) Fourier analysis of the attenuated Radon transform and a 2-D version of the Shannon sampling theorem, we investigate the problem of resolution in dynamic emission tomography. As a result we provide guidelines on how to acquire and on how to filter the projection data.

Key words. dynamic emission tomography, photon transport equation, attenuated Radon transform, 2-D sampling theorem, resolution, aliasing errors, uncertainty principle, 2-D filtering

AMS subject classifications. 94A20, 44A12, 92C55

PII. S0036141098345457

1. Introduction.

Current state-of-the-art medical imaging technologies provide extremely detailed and accurate information about human anatomy. However, the corresponding detailed information about function is not yet readily available. *Single photon emission computerized tomography* (SPECT) is a noninvasive diagnostic technology which is used to show the blood flow in the heart muscle, extent of damage in stroke patients, presence and degree of malignancy of tumors, and much else.

SPECT is able to image the function of the body through a *tracer*, a biochemical molecule labeled with radioactivity. The radioactive material is incorporated by the patient and metabolized by the organ of interest. The emissions are then recorded by a rotating SPECT camera (cf. Figure 1), and a three-dimensional (3-D) visualization is created from the two-dimensional (2-D) projection data.

Currently, the data recorded by SPECT cameras are static and qualitative. It is not possible, as yet, to measure absolute metabolic rates from the different biological processes, nor to measure the movement of molecules during biodistribution and metabolism. Recently, a major step toward the development of *dynamic* SPECT (dSPECT) has been achieved through two mathematical methods replacing the traditional filtered backprojection (FBP) method (cf. [2, 14, 16, 5]), the latter being by its nature static (cf. [18]) and not feasible for dynamic sources.

The present paper will focus on the problem of resolution in dynamic emission tomography. Results of this type have previously been obtained in static SPECT, in positron emission tomography (PET), and in computed tomography (CT), where an elaborate Fourier analysis led to the idea, among others, of *interlaced grids* which significantly improved resolution (cf. [13, 19, 20, 22]). Following these lines, we shall answer typical questions like how many positions a SPECT camera should take, how long it should stay in a given position, whether views should be recorded over 180 degrees or 360 degrees, or what the internal resolution of the camera should be if a certain spatial resolution in the reconstructed image has to be achieved. As a second application, we present some ideas on how to filter data before doing the actual inversion.

*Received by the editors October 5, 1998; accepted for publication (in revised form) June 23, 1999; published electronically May 2, 2000.
http://www.siam.org/journals/sima/31-5/34545.html

†Mathématiques pour l'Industrie et la Physique, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France (maeght@mip.ups-tlse.fr, noll@dumbo.ups-tlse.fr).

2. The model. Emission tomography is modeled by the 3-D dynamic photon transport equation (cf. [8]). It is convenient to simplify the model by assuming that scattering is negligible or, rather, to interpret it as a measurement noise. This decouples the equation and allows for splitting the 3-D reconstruction into a series of 2-D reconstructions on slices. The simplified dynamic 2-D transport equation is

$$\frac{1}{c}\,u_t(t,x,\omega,E)+\omega\cdot\nabla u(t,x,\omega,E)+\mu(x,E)\,u(t,x,\omega,E)=f(t,x,E), \tag{2.1}$$

where $u(t,x,\omega,E)$ is the (unknown) photon transport at time t and position $x\in\mathbf{R}^2$ at the energy level E in direction $\omega\in\mathbf{S}^1$, $\mu(x,E)$ is the unknown linear attenuation coefficient at position x for photons traveling with energy E, and $f(t,x,E)$ is the unknown number of photons emitted at time t, position x, at energy level E.

As opposed to X-rays, γ-rays are monochromatic, i.e., photons are emitted at a fixed energy level E_0, for instance, $E_0 = 140$keV for Technetium used in many clinical applications. Similarly, in PET, the recorded photons, originating from the annihilation of a positron with an electron, travel with $E_0 = 511$keV, the energy of the electron. Photons recorded with energy $E < E_0$ are therefore due to Compton scatter, and an energy window $\pm\Delta E$ about the expected level E_0 allows for eliminating most scattering events (cf. [25]). It is legitimate to further simplify (2.1) by omitting the reference to energy. More precisely, writing $f(t,x,E)=f(t,x)\,\delta(E-E_0)$ and $\mu(x)=\mu(x,E_0)$, the equation for the cumulative transport $u(t,x,\omega)$ integrated over the relevant energy levels $E\in[E_0-\Delta E,E_0+\Delta E]$ is

$$\frac{1}{c}\,u_t(t,x,\omega)+\omega\cdot\nabla u(t,x,\omega)+\mu(x)\,u(t,x,\omega)=f(t,x), \tag{2.2}$$

which may be solved explicitly on each line.

To do this, we have to supply boundary conditions. We assume that the unknown source and attenuation coefficient are supported on the unit disk D. We adopt the notations $\omega=(\cos\phi,\sin\phi)$ and $\omega^\perp=(-\sin\phi,\cos\phi)$. Rays may then be referenced (s,ϕ), that is, $x\cdot\omega^\perp=s$, or $x=s\omega^\perp+\tau\omega$, $\tau\geq 0$ for x on the ray so referenced. Now notice that the incoming radiation is zero, i.e.,

$$u(t,s\omega^\perp+\tau\omega,\omega)=0\quad\text{for all }\ \tau\leq\tau_0=\tau_0(s,\omega)\ \text{ and all } t$$

($x_0=s\omega^\perp+\tau_0\omega$ the entry point of the ray $x\cdot\omega^\perp=s$ into D, if any). Second, we use the fact that $u(t,x,\omega)$ has been recorded at certain times t and for certain directions ω at a camera bin located at $x_1=s\omega^\perp+\tau_1\omega$, $\tau_1=\tau_1(s,\omega)$, on the line $x\cdot\omega^\perp=s$. (Without loss, we may assume that x_1 is the exit point of the ray $x\cdot\omega^\perp$ from the disk D, if any.) That is, the observed data are of the form

$$u(t,s\omega^\perp+\tau_1\omega,\omega)=:d(s,\omega,t).$$

In fact, a SPECT camera (shown schematically in Figure 1) detects photons which arrive perpendicular to the camera surface while at a fixed angular position $\omega=(\cos\phi,\sin\phi)$.

Integrating (2.2) using the boundary conditions gives the nonlinear relation

$$\int_{\tau_0}^{\tau_1} f(t+(\tau_1-\tau)/c,s\omega^\perp+\tau\omega)\,e^{-\displaystyle\int_{\tau}^{\tau_1}\mu(s\omega^\perp+\rho\omega)\,d\rho}\,d\tau\;=\;d(s,\omega,t). \tag{2.3}$$

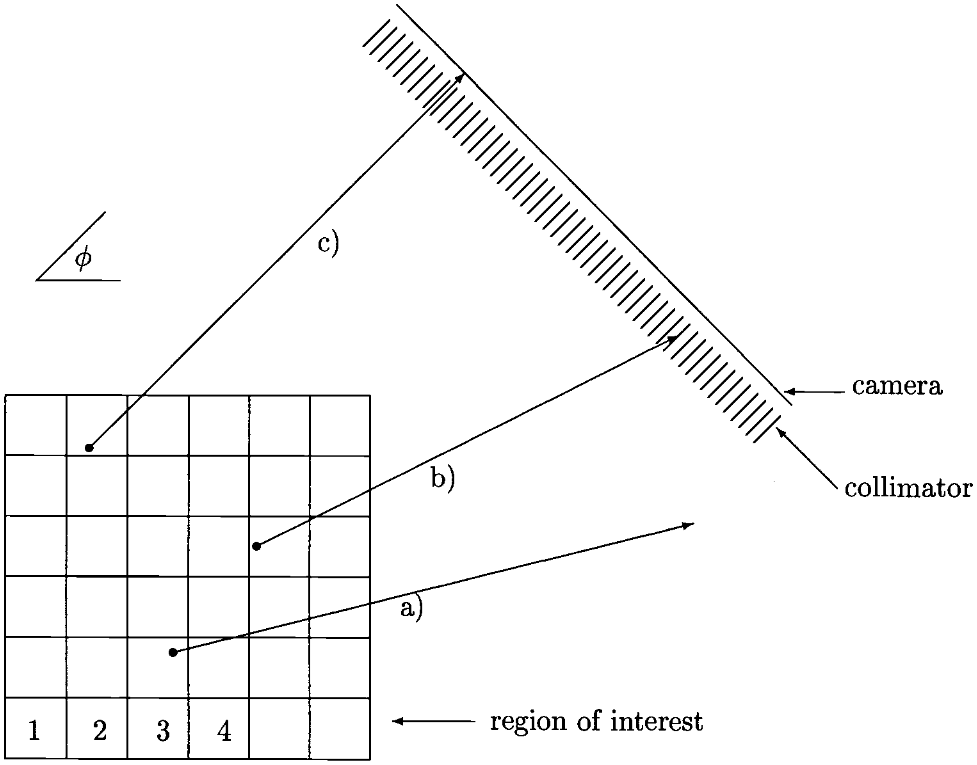


FIG. 1. *The principle of SPECT: Photons radiating from the region of interest.* (a) *Photon misses the camera,* (b) *is absorbed by the collimator,* (c) *passes the collimator and hits the camera. The camera rotates around the region of interest (schematically discretized into pixels).*

As photons are traveling with the speed of light c, in practice $t+(\tau_1-\tau_0)/c \approx t$, and (2.3) simplifies to

$$(2.4)\ R[\mu, f(t,\cdot)](s,\omega) := \int_{\tau_0}^{\tau_1} f(t, s\omega^{\perp}+\tau\omega)\, e^{-\int_{\tau}^{\tau_1} \mu(s\omega^{\perp}+\rho\omega)\,d\rho}\, d\tau \;=\; d(s,\omega,t).$$

Here $R[\mu, f]$ denotes the *attenuated Radon transform* (cf. [18, 20, 21]). Solving (2.4) simultaneously for the unknown dynamic source $f(t,x)$ and attenuation $\mu(x)$, based on the acquired data $d(s,\omega,t)$, is the mathematical problem of dSPECT. A similar equation replacing (2.4) may be found for PET using the corresponding symmetric data (see [16]).

Attempts to estimate *both* the unknown attenuation and source term from the projection data have been made by several authors. A method proposed by Natterer [18, 21] and reported to be practical in [24] uses a consistency condition to obtain an estimate of μ. This approach is feasible if the camera takes views over 360 degrees and the source is static. More recently, Dicken [9] proposed a direct inversion of (2.4). In practice, often less sophisticated ways are chosen, which consist either in neglecting attenuation or assuming constant attenuation (after tracing the contour of the patient) or in correcting data via some heuristic methods. A third way proposed in [6, 7] consists in doing a CT in parallel with the emission scan. The remaining problem of estimating f with known μ is then linear but still ill-posed (cf. [20]).

3. Fourier analysis. The way the inverse problem (2.4) is solved depends on the device. A ring SPECT camera, and similarly the PET cameras, allow for collecting a full set of angular views $d(s_j, \omega_k, t_\ell)$, (j the index for camera bins, k the index for angular positions, ℓ the index for stops) at a fixed time t_ℓ. The reconstruction algorithms are then essentially the static ones, FBP or EM algorithms, which reconstruct one static image at a time t_ℓ, obtaining the dynamic image frame by frame. The 2-D Fourier analysis of the static case being known, cf. [20, 13, 22], we may consider this case as essentially understood.

The situation changes if a rotating camera system is used. As the activity in the organ changes significantly during the scan, a rotating camera (even triple head) will not be able to collect sufficiently many views ω_k at a fixed time t_ℓ in order to reconstruct the dynamic object frame by frame. Rather, in the extreme case of a single head camera, we can scan only one position at a time, so the acquired data are $d(s_j, \omega_k, t_k)$. Ideally, the time axis t and angular position ϕ are then linked through

$$t = \frac{T}{2\pi}\,\phi, \qquad 0 \le t \le T,\ 0 \le \phi \le 2\pi \tag{3.1}$$

(T is the total acquisition time). With (3.1), reconstruction algorithms necessarily have to process all the projection data simultaneously, which leads to large size problems difficult to solve in practice (cf. [2, 16], and also [12, 5, 10, 14, 17]).

Assuming that the dynamic source is of the form $f(t,x) = g_1(t)\,h_1(x) + \cdots + g_r(t)\,h_r(x)$, its attenuated Radon transform (2.4) is

$$R[\mu, f(t,\cdot)](s,\phi) = g_1(t)\,R[\mu, h_1](s,\phi) + \cdots + g_r(t)\,R[\mu, h_r](s,\phi).$$

In the case of a rotating camera, in particular a single head camera, (3.1) leads to the ideal projection data

$$p(s,\phi) := g_1((T/2\pi)\phi)\,R[\mu, h_1](s,\phi) + \cdots + g_r((T/2\pi)\phi)\,R[\mu, h_r](s,\phi), \tag{3.2}$$

often referred to as the *sinogram* of the source $f(t,x)$, for the obvious reason that a point source scanned over 360 degrees would produce a sinoidal curve. Figures 4(c) and 4(g) show some experimental sinogram data collected over a 180-degree scan.

The principal purpose of the present paper is to perform a 2-D Fourier analysis of the sinogram $p(s,\phi)$. As a result of this analysis we obtain two practical guidelines:

(1) On *resolution.* How many stops and angular positions are required to capture a prescribed spatial resolution along with a predicted half-life? How long should an individual stop last?

(2) On *data filtering*, which is inherent to the classical FBP algorithms but has to be considered anew in dSPECT.

4. Sampling in two-dimensions. In this section, we shall be concerned with the sampling of the sinogram (3.2) of a dynamic source $f(t,x)$. In the first round we shall consider only the unattenuated Radon transform (i.e., $\mu = 0$). Later on we will indicate that the results are usually not altered if attenuation is taken into account.

We recall that the Radon transform $Rh(s,\phi)$ of a spatial function $h(x)$, being 2π-periodic in ϕ, is defined on $\mathbf{R} \times \mathbf{S}^1$, which we shall call the (s,ϕ)-plane or *physical plane.* The 2-D Fourier transform $\hat{p}$ of $p(s,\phi)$ is then defined on $\mathbf{R} \times \mathbf{Z}$, which will be referred to as the (σ, k)-plane or *frequency plane.*

A sampling operator $\mathcal{S}_{K,W}$ in the physical plane is defined by two ingredients—a *sampling lattice* $W\mathbf{Z}^2$ in the physical plane (W a 2×2-matrix) in tandem with a *spectral window* K in the frequency plane—whose replica $K + 2\pi(W^{-1})^T\ell$, $\ell \in \mathbf{Z}^2$, generated by the *dual lattice* $2\pi(W^{-1})^T\mathbf{Z}^2$ in the frequency-plane, are mutually disjoint:

$$\mathcal{S}_{K,W}p(s,\phi) := \det(W) \sum_{\ell\in\mathbf{Z}^2} p(W\ell)\,\hat{\chi}_K((s,\phi) - W\ell) \tag{4.1}$$

(χ_K the characteristic function of the set K). More formally, $\mathcal{S}_{K,W}$ may be represented using the shah-distribution $\text{Ш}(s,\phi) = \sum_{\ell\in\mathbf{Z}^2} \delta(s-\ell_1, \phi-\ell_2)$,

$$\mathcal{S}_{K,W}p = \Big(p \cdot \text{Ш}(W^{-1}\cdot)\Big) * \hat{\chi}_K,$$

a formulation which is very intuitive when we consider its Fourier transform. Replacing the analog signal p by its digitized version $p \cdot \text{Ш}(W^{-1}\cdot)$, taken at the points of the lattice $W\mathbf{Z}^2$, has the following effect: Since $\hat{\text{Ш}} = \text{Ш}$, the spectrum of the digitized signal shows the true spectrum, $\hat{p}$, but repeated periodically

$$\widehat{\big(p \cdot \text{Ш}(W^{-1}\cdot)\big)} = 2\pi \sum_{\ell\in\mathbf{Z}^2} \hat{p}(\cdot - 2\pi W^{-T}\ell)$$

along the dual lattice $2\pi W^{-T}\mathbf{Z}^2$. Consequently, if the spectral window K is well chosen, i.e., if the spectrum $\hat{p}$ is captured by K, we may fully retrieve the true signal p, simply by applying an ideal low pass filter χ_K which eliminates frequencies $\notin K$:

$$\widehat{\big(\mathcal{S}_{K,W}p\big)} = (2\pi)^{-1}\widehat{\big(p \cdot \text{Ш}(W^{-1}\cdot)\big)} \cdot \chi_K = \sum_{\ell\in\mathbf{Z}^2} \hat{p}(\cdot - 2\pi W^{-T}\ell)\cdot\chi_K. \tag{4.2}$$

In fact, 2-D versions of the Shannon sampling theorem are easily understood through (4.2): the signal p is fully retrieved from the sampled signal if its spectrum has $\text{supp}(\hat{p}) \subset K$. In our applications, however, we are dealing with compactly supported signals, whose spectra $\hat{p}$ are analytic and never fully supported on a bounded set K. We will consequently have to accept aliasing errors associated with the choice of a sampling operator (4.1). Estimating these errors is the principal task of the present section. Practical aspects will be considered later.

Let us fix $0 < \vartheta < 1$, a positive integer m, and $b > 0$. As our frequency window in the (σ, k)-plane we choose the *bowtie region* K,

$$K = \{(\sigma,k) \in \mathbf{R}\times\mathbf{Z} : |\sigma| \le b \text{ and } |k| \le |\sigma|/\vartheta + m\}, \tag{4.3}$$

which is displayed in Figure 2(a). Figure 2(c) indicates the scheme $2\pi W^{-T}\mathbf{Z}^2$ which produces nonoverlapping replica of K in the frequency plane. The sampling parameters are seen to be $\Delta k = [b/\vartheta] + 2m$ and $\Delta\sigma = b$, and the matrices W, $2\pi W^{-T}$ are

$$W = 2\pi\begin{pmatrix} \frac{1}{2b} & 0 \\ -\frac{1}{2\Delta k} & \frac{1}{\Delta k}\end{pmatrix}, \quad 2\pi(W^{-1})^T = \begin{pmatrix} 2b & b \\ 0 & \Delta k\end{pmatrix}. \tag{4.4}$$

As we shall see in our experiments, the parameter ϑ may in practice be chosen as $\vartheta \approx 1$ but for theoretical reasons has to satisfy $\vartheta \in (0,1)$.

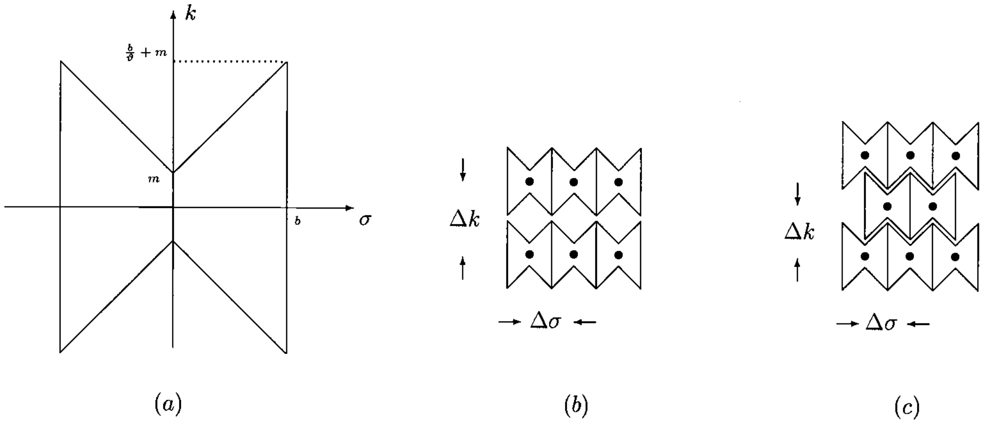


FIG. 2. (a) *shows the bowtie region K defined through* (4.3), (4.4), *while* (b) *and* (c) *show two different lattices generating disjoint replica of K. The interlaced grid* (c) *requires fewer nodes and therefore gives a better sampling scheme.*

Naturally, the identification of the essential support K of the spectrum $\hat{p}$ in tandem with the sampling lattice $W\mathbf{Z}^2$ is crucial for our principal tasks: resolution and filtering. The choice of K being ambiguous, we shall have to support our proposition (4.3) both by numerical tests and by a rigorous analysis, including error estimates. In the present section we proceed to provide those.

To formulate our main result, we need to introduce two notions for the spatial and temporal bandwidths, respectively. Concerning the spatial term, Natterer [20] considers the measures

$$\epsilon_d(h,b) = \int_{|\xi|>b} |\xi|^d |\hat{h}(\xi)|\, d\xi,$$

which may be related to appropriate Sobolev norms. In fact, using

$$\|h\|_{W^{\alpha,d}} = \left(\int_{\mathbf{R}^n} (1+|\xi|^2)^{d\alpha/2}\, |\hat{h}(\xi)|^{\alpha}\, d\xi \right)^{1/\alpha},$$

and defining the ideal *high pass filter at frequency* b, $\mathcal{H}_b$, via $\big(\mathcal{H}_b h\big)\hat{} = \hat{h}\chi_{\{|\cdot|>b\}}$, Natterer's error terms satisfy $C_1\epsilon_d(h,b) \le \|\mathcal{H}_b h\|_{W^{1,d}} \le C_2\epsilon_d(h,b)$. For the following we shall, in addition, obtain error estimates involving the Hilbert space norms $\|\mathcal{H}_b h\|_{W^{2,d}}$.

In turn, for a 2π-periodic function $g(\phi)$ with Fourier coefficients $\hat{g}_k$, low pass filtering at a frequency k obviously corresponds to truncating the Fourier series at k, and correspondingly, high pass filtering corresponds to retaining the tail $|\nu| > k$ of the series. We therefore consider the error terms

$$R_k(g) := \left(\sum_{|\nu|>k} |\hat{g}_\nu|^2 \right)^{1/2},$$

$k > 0$, which play a role similar to the norm estimates of $\mathcal{H}_b h$ above. With these notions we are ready to state the following theorem.

THEOREM 4.1. *Let ϑ, m, b and K, W be as in* (4.3), (4.4). *Choose $\vartheta' \in (\vartheta, 1)$, and let $\theta := \vartheta/\vartheta' \in (0,1)$. Consider a dynamic source of the form $f(t,x) = g(t)h(x)$ such* *that $h(x)$ is continuous and supported on the unit disk D and $g(t)$ is continuous. Let $g((T/2\pi)\phi)$ be continued periodically for $\phi \notin [0, 2\pi]$. Let $p(s,\phi)$ be the ideal sinogram of $f(t,x)$, and let $\mathcal{S}_{K,W}p(s,\phi)$ be the sinogram sampled on the lattice $W\mathbf{Z}^2$ in the (s,ϕ)-plane using the frequency window K. Then*

$$\begin{aligned}
1.\ \|p - \mathcal{S}_{K,W}p\|_\infty &= \|g\|_\infty \mathcal{O}\Big(m\epsilon_{-1}(h,b) + \epsilon_0(h,b)\Big) \\
&\quad + \|h\|_\infty \mathcal{O}\Big(\sum_{\nu=1}^{\infty} \nu R_{(1-\vartheta')(\nu+m)}(g)\Big), \\
2.\ \|p - \mathcal{S}_{K,W}p\|_2 &= \|g\|_\infty \mathcal{O}\Big(\|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}}\Big) \\
&\quad + \|h\|_\infty \mathcal{O}\Big(\Big(\sum_{\nu=1}^{\infty} \nu R_{(1-\vartheta')(\nu+m)}(g)^2\Big)^{1/2}\Big), \\
3.\ \|p - \mathcal{S}_{K,W}p\|_2 &= \|g\|_\infty \mathcal{O}\Big(\|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}}\Big) \\
&\quad + \|h\|_\infty \mathcal{O}\Big(b^{1/2}\Big(\sum_{\nu=1}^{\infty} R_{(1-\vartheta')(\nu+m)}(g)^2\Big)^{1/2}\Big).
\end{aligned}$$

Proof. Part 1. Notice that by Parseval's formula, $\|p\|_2 = \|\hat{p}\|_2$, where $\hat{p}$ is the 2-D Fourier transform of p, and according to [20, p. 63], $\|p\|_\infty \le \|\hat{p}\|_1$, if the corresponding norms on the frequency plane are defined through

$$\|\hat{p}\|_\alpha = \Big(\sum_{k=-\infty}^{\infty} \int_{\mathbf{R}} |\hat{p}(\sigma,k)|^\alpha \, d\sigma \Big)^{1/\alpha}. \tag{4.5}$$

By the definition of the sampling operator (4.1),

$$\hat{p} - (\mathcal{S}_{K,W}p)\hat{} = (1-\chi_K)\,\hat{p} - \sum_{\ell \ne 0} \hat{p}(\cdot - 2\pi W^{-T}\ell)\,\chi_K,$$

so we derive the estimate

$$\|(p - \mathcal{S}_{K,W}p)\hat{}\|_\alpha \le \|(1-\chi_K)\hat{p}\|_\alpha + \Big\| \sum_{\ell \ne 0} \hat{p}(\cdot - 2\pi W^{-T}\ell)\chi_K \Big\|_\alpha .$$

By the translation invariance of the Haar measure and using the fact that the translates $K + 2\pi W^{-T}\ell$ are disjoint, the second term on the right-hand side satisfies

$$\Big\| \sum_{\ell \ne 0} \hat{p}(\cdot - 2\pi W^{-T}\ell)\chi_K \Big\|_\alpha \le \|(1-\chi_K)\hat{p}\|_\alpha,$$

so all together

$$\|(p - \mathcal{S}_{K,W}p)\hat{}\|_\alpha \le 2\|(1-\chi_K)\hat{p}\|_\alpha.$$

Writing $K(k) = \{\sigma : (\sigma,k) \in K\}$, we are therefore led to estimate

$$\sum_{k=-\infty}^{\infty} \int_{\sigma \notin K(k)} |\hat{p}(\sigma,k)|^\alpha \, d\sigma. \tag{4.6}$$

To do this, we will have to distinguish the cases $\alpha = 1$ and $\alpha = 2$. For the case $\alpha = 1$, we decompose the region $(\sigma, k) \notin K$ into three parts Σ_1, Σ_2, and Σ_3:

$$\begin{aligned}
\Sigma_1 &= \{(\sigma, k) : |k| > |\sigma|/\vartheta + m\},\\
\Sigma_2 &= \{(\sigma, k) : |\sigma| > b \text{ and } |k| \leq b/\vartheta + m\},\\
\Sigma_3 &= \{(\sigma, k) : |\sigma| > b \text{ and } |k| \leq |\sigma|/\vartheta + m \text{ and } |k| \geq b/\vartheta + m\}.
\end{aligned}$$

For the case $\alpha = 2$, two domains Γ_1, Γ_2 will do:

$$\begin{aligned}
\Gamma_1 &= \{(\sigma, k) : |k| \geq m \text{ and } |\sigma| \leq \min\big(b, \vartheta(|k| - m)\big)\},\\
\Gamma_2 &= \{(\sigma, k) : |\sigma| > b\}.
\end{aligned}$$

Part 2. Let us continue collecting useful information. Let $\mathcal{F}_s p(\cdot, \phi)(\sigma)$ be the one-dimensional (1-D) Fourier transform of p with respect to s, which is again 2π-periodic in ϕ. Its kth Fourier coefficient is

$$\hat{p}(\sigma, k) =: \hat{p}_k(\sigma) = \frac{1}{2\pi} \int_0^{2\pi} \mathcal{F}_s p(\sigma, \phi) e^{-ik\phi}\, d\phi.$$

By the Fourier slice theorem [20, p. 11] we have

$$\mathcal{F}_s p(\sigma, \phi) = g((T/2\pi)\phi)\, \mathcal{F}_s[Rh(\cdot, \phi)](\sigma) = g((T/2\pi)\phi)\, \hat{h}(\sigma\omega)$$

with $\omega = (\cos\phi, \sin\phi)$. Then

$$\begin{aligned}
\hat{p}_k(\sigma) &= (2\pi)^{-1/2} \int_0^{2\pi} g((T/2\pi)\phi) \hat{h}(\sigma\omega) e^{-ik\phi}\, d\phi \\
&= (2\pi)^{-3/2} \int_0^{2\pi} g((T/2\pi)\phi) \int_D e^{-i\sigma\omega\cdot x} h(x)\, dx\ e^{-ik\phi}\, d\phi \\
&= (2\pi)^{-3/2} \int_D h(x) \int_0^{2\pi} g((T/2\pi)\phi) e^{-i\sigma|x|cos(\phi-\psi)-ik\phi}\, d\phi\, dx
\end{aligned} \tag{4.7}$$

if we put $x = |x|(\cos\psi, \sin\psi)$. Substituting the Fourier series $g((T/2\pi)\phi) = \sum_\nu \hat{g}_\nu e^{i\nu\phi}$, this becomes

$$\begin{aligned}
\hat{p}_k(\sigma) &= (2\pi)^{-3/2} \int_D h(x) e^{-ik\psi} \sum_{\nu=-\infty}^{\infty} \hat{g}_\nu \int_0^{2\pi} e^{-i\sigma|x|\cos\phi - i(k-\nu)\phi}\, d\phi\, dx \\
&= (2\pi)^{-1/2} i^k \int_D h(x) e^{-ik\psi} \sum_{\nu=-\infty}^{\infty} \hat{g}_\nu J_{k-\nu}(-\sigma|x|)\, dx, \qquad (4.8)
\end{aligned}$$

where we have used the Bessel functions $J_k(x)$ of the first kind defined through

$$J_k(x) = \frac{i^{-k}}{2\pi} \int_0^{2\pi} e^{ix\cos\phi - ik\phi}\, d\phi.$$

Part 3. Let us now consider the estimate on Σ_2 with $\alpha = 1$. As a consequence of (4.7),

$$\begin{aligned}
\int_{|\sigma|>b} |\hat{p}_k(\sigma)|\, d\sigma &\leq (2\pi)^{-1/2} \|g\|_\infty \int_{|\sigma|>b} \int_0^{2\pi} |\hat{h}(\sigma\omega)|\, d\phi\, d\sigma \\
&= (2\pi)^{-1/2} \|g\|_\infty \int_{|\xi|>b} \frac{|\hat{h}(\xi)|}{|\xi|}\, d\xi = (2\pi)^{-1/2} \|g\|_\infty \epsilon_{-1}(h, b). \qquad (4.9)
\end{aligned}$$

Now observe that $(\sigma, k) \in \Sigma_2$ only for $|k| \leq b/\vartheta + m$, so we are left with a finite number of terms (4.9). In fact

$$\begin{aligned}\|\chi_{\Sigma_2}\hat{p}\|_1 &= \sum_{|k|\leq b/\vartheta+m} \int_{|\sigma|>b} |\hat{p}_k(\sigma)|\, d\sigma \leq C\|g\|_\infty (2b/\vartheta + 2m + 1)\epsilon_{-1}(h, b) \\ &\leq C\|g\|_\infty \big(b\epsilon_{-1}(h, b) + m\epsilon_{-1}(h, b)\big) \leq C\|g\|_\infty \big(\epsilon_0(h, b) + m\epsilon_{-1}(h, b)\big), \qquad (4.10)\end{aligned}$$

where the last inequality uses (2.9) in [20, p. 66].

Part 4. Let us next discuss the estimate on Σ_3, $\alpha = 1$. According to (4.9) above, replacing b by $\vartheta(|k| - m)$ gives

$$\int_{|\sigma|\geq\vartheta(|k|-m)} |\hat{p}_k(\sigma)|\, d\sigma \leq C\|g\|_\infty \epsilon_{-1}\big(h, \vartheta(|k| - m)\big).$$

But then

$$\begin{aligned}\|\chi_{\Sigma_3}\hat{p}\|_1 &= \sum_{|k|\geq b/\vartheta+m} \int_{|\sigma|\geq\vartheta(|k|-m)} |\hat{p}_k(\sigma)|\, d\sigma \\ &\leq C\|g\|_\infty \sum_{|k|\geq b/\vartheta+m} \epsilon_{-1}(h, \vartheta(|k| - m)) \leq C\|g\|_\infty \epsilon_0(h, b), \qquad (4.11)\end{aligned}$$

where the last inequality again uses (2.9) in [20, p. 66].

Part 5. Let us catch up with the error estimate on Γ_2, $\alpha = 2$. Using (4.7), for fixed σ, and with $\omega = (\cos\phi, \sin\phi)$, Parseval's equality gives

$$\sum_{k=-\infty}^{\infty} |\hat{p}_k(\sigma)|^2 = \int_0^{2\pi} |g((T/2\pi)\phi)\, \hat{h}(\sigma\omega)|^2\, d\phi.$$

Integrating over $|\sigma| > b$ shows

$$\begin{aligned}\|\chi_{\Gamma_2}\hat{p}\|_2^2 &\leq C\|g\|_\infty^2 \int_{|\sigma|>b} \int_0^{2\pi} |\hat{h}(\sigma\omega)|^2 d\phi\, d\sigma \\ &= C\|g\|_\infty^2 \int_{|\xi|>b} \frac{|\hat{h}(\xi)|^2}{|\xi|}\, d\xi \leq C\|g\|_\infty^2 \|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}}^2. \qquad (4.12)\end{aligned}$$

Part 6. We need to consider some technical preliminaries about Bessel functions. It is well known that $J_n(\sigma)$ decays exponentially ($n \to \infty$) for fixed σ. According to [23] even $J_n(\theta n) \to 0$ exponentially, since $0 < \theta < 1$. We'll improve this by showing that, regarding $0 < \theta < 1$, $R_n(J_\cdot(\theta n)) \to 0$ exponentially as $n \to \infty$.

According to [23, p. 255] in tandem with [1, Theorem 4.1.28], and regarding $0 < \theta < 1$, we have (for n a positive integer)

$$0 \leq J_n(\theta n) \leq (2\pi n)^{-1/2}(1 - \theta^2)^{-1/4} e^{-(n/3)(1-\theta^2)^{3/2}}.$$

Summing over $|\nu| \geq n$ and using $|J_\nu(-\sigma)| = |J_\nu(\sigma)|$, $|J_{-\nu}(\sigma)| = |J_\nu(\sigma)|$ gives

$$\begin{aligned}\sum_{|\nu|\geq n} |J_\nu(\theta n)|^2 &= 2\sum_{\nu=n}^{\infty} \big|J_\nu\big((\theta n/\nu)\cdot\nu\big)\big|^2 \\ &\leq 2\sum_{\nu=n}^{\infty} \left((2\pi\nu)^{-1/2}(1 - (\theta n/\nu)^2)^{-1/4} e^{-(\nu/3)(1-(\theta n/\nu)^2)^{3/2}}\right)^2.\end{aligned}$$

This term is easily seen to decay exponentially in n, that is,

$$|R_n(J.(\theta' n))| \le Ce^{-c(\theta)n} \tag{4.13}$$

for a constant $c(\theta) > 0$ depending on θ, and uniformly over $0 \le \theta' \le \theta$.

Part 7. Let us now consider the error estimates on Σ_1 with $\alpha = 1$. Writing

$$d_k(\sigma) := \sum_{\nu=-\infty}^{\infty} \hat{g}_\nu J_{k-\nu}(\sigma) = \big(\hat{g} * J.(\sigma)\big)_k \qquad \text{(convolution of sequences)},$$

we find using (4.8) that

$$|\hat{p}_k(\sigma)| \;\le\; C\|h\|_\infty \max_{|x|\le 1} d_k(-\sigma|x|), \tag{4.14}$$

and by the definition of Σ_1, we are led to estimate $d_k(-\sigma|x|)$ for $|\sigma| \le \vartheta(|k| - m)$, $|k| \ge m$, and $|x| \le 1$. This is done by the following steps. Observe that

$$\begin{aligned} |d_k(-\sigma|x|)| &\le \sum_{\nu=-\infty}^{\infty} |\hat{g}_\nu J_{k-\nu}(-\sigma|x|)| \\ &= \sum_{|\nu|\ge(1-\vartheta')|k|} |\hat{g}_\nu J_{k-\nu}(-\sigma|x|)| + \sum_{|\nu|<(1-\vartheta')|k|} |\hat{g}_\nu J_{k-\nu}(-\sigma|x|)| =: \mathrm{I} + \mathrm{II}. \end{aligned}$$

The first term I satisfies

$$\mathrm{I} \le \Bigg(\sum_{|\nu|\ge(1-\vartheta')|k|} |\hat{g}_\nu|^2 \Bigg)^{1/2} \Bigg(\sum_{|\nu|\ge(1-\vartheta')|k|} |J_{k-\nu}(-\sigma|x|)|^2 \Bigg)^{1/2} \le R_{(1-\vartheta')|k|}(g), \tag{4.15}$$

the second factor being ≤ 1 since $\sum_k |J_k(z)|^2 = 1$ for every z. The second term II is estimated through

$$\mathrm{II} \le \|g\|_2 \Bigg(\sum_{|\nu|<(1-\vartheta')|k|} |J_{k-\nu}(-\sigma|x|)|^2 \Bigg)^{1/2} \le \|g\|_2 R_{\vartheta'|k|}\big(J.(-\sigma|x|)\big). \tag{4.16}$$

Here the argument $-\sigma|x|$ of the Bessel coefficients satisfies

$$\big| -\sigma|x|\big| \le |\sigma| \le \vartheta(|k| - m) = \vartheta'|k| \, \frac{\vartheta(|k|-m)}{\vartheta'|k|} \le \vartheta'|k| \cdot \frac{\vartheta}{\vartheta'} = \vartheta'|k| \cdot \theta,$$

which means that $| -\sigma|x|| = \theta'\vartheta'|k|$ for some $0 \le \theta' \le \theta$. The latter allows us to apply (4.13) with $n = \vartheta'|k|$:

$$R_{\vartheta'|k|}(J.(-\sigma|x|)) \;\le\; Ce^{-\gamma(\theta)|k|} \qquad \text{for } \ |\sigma| \le \vartheta(|k| - m) \tag{4.17}$$

and some $\gamma(\theta) > 0$.

Finally, using in this order (4.14), (4.15), (4.16), and (4.17), the error term on Σ_1 is

$$\begin{aligned} \|\chi_{\Sigma_1}\hat{p}\|_1 &\le C\|h\|_\infty \max_{|x|\le 1} \sum_{|k|\ge m} \int_{|\sigma|\le\vartheta(|k|-m)} |d_k(-\sigma|x|)| \, d\sigma \\ &\le C\|h\|_\infty \sum_{|k|\ge m} (|k| - m) \Big(R_{(1-\vartheta')|k|}(g) + \|g\|_2 e^{-\gamma(\theta)|k|} \Big) \\ &\le C\|h\|_\infty \Bigg(\sum_{\nu=1}^{\infty} \nu R_{(1-\vartheta')(\nu+m)}(g) + \|g\|_2 e^{-\delta(\theta)m} \Bigg) \end{aligned} \tag{4.18}$$

for another constant $\delta(\theta) > 0$. Clearly then, (4.18) shows us that $\|\chi_{\Sigma_1}\hat{p}\|_1 = \|h\|_\infty \mathcal{O}(\sum_{\nu=1}^\infty \nu R_{(1-\vartheta')(\nu+m)}(g))$, the exponentially decaying term being negligible. Combining this with (4.10) and (4.11) gives statement 1.

Part 8. Our last step is the error estimate $\|\chi_{\Gamma 1}\hat{p}\|_2$. Notice that by the definition of Γ_1,

$$\begin{aligned}\|\chi_{\Gamma_1}\hat{p}\|_2^2 &\le \sum_{|k|\ge m} \int_{|\sigma|\le \min\{b,\vartheta(|k|-m)\}} |\hat{p}_k(\sigma)|^2\, d\sigma \\ &\le \sum_{m\le |k| \le \frac{b}{\vartheta}+m} \int_{|\sigma|\le\vartheta(|k|-m)} |\hat{p}_k(\sigma)|^2\, d\sigma + \sum_{|k|\ge \frac{b}{\vartheta}+m} \int_{|\sigma|\le b} |\hat{p}_k(\sigma)|^2\, d\sigma =: \mathrm{III}^2 + \mathrm{IV}^2.\end{aligned}$$

Using (4.14) and (4.17), we find

$$\begin{aligned}\mathrm{III}^2 &\le C\|h\|_\infty^2 \max_{|x|\le 1} \sum_{m\le |k|\le \frac{b}{\vartheta}+m} \int_{|\sigma|\le\vartheta(|k|-m)} |d_k(-\sigma|x|)|^2\, d\sigma \\ &\le C\|h\|_\infty^2 \sum_{m\le |k|\le \frac{b}{\vartheta}+m} \int_{|\sigma|\le\vartheta(|k|-m)} \left(R_{(1-\vartheta')|k|}(g) + \|g\|_2 e^{-\delta|k|}\right)^2 d\sigma,\end{aligned}$$

(4.17) being applicable since $|-\sigma|x|| \le \vartheta(|k|-m) < \vartheta'|k|$. By the triangle inequality, and on setting $\nu = |k|-m$,

$$\mathrm{III} \le C\|h\|_\infty \left(\left(\sum_{\nu=1}^{b/\vartheta} \nu R_{(1-\vartheta')(\nu+m)}(g)^2\right)^{1/2} + \|g\|_2 e^{-\gamma m}\right)$$

for another constant $\gamma > 0$. Similarly, the term IV satisfies

$$\mathrm{IV} \le C\|h\|_\infty \left(b^{1/2}\left(\sum_{\nu=b/\vartheta}^{\infty} R_{(1-\vartheta')(\nu+m)}(g)^2\right)^{1/2} + \|g\|_2 e^{-\delta m}\right).$$

So all together,

$$\|\chi_{\Gamma_1}\hat{p}\|_2 \le C\|h\|_\infty \left(\left(\sum_{\nu=1}^{\infty} \min\{\nu,b\} R_{(1-\vartheta')(\nu+m)}(g)^2\right)^{1/2} + \|g\|_2 e^{-\delta m}\right),$$

which in tandem with (4.12) responds to estimates 2 and 3, again since the exponentially decaying term is negligible. □

The function $h(x)$, supported on the unit disk D, is called *essentially bandlimited* if

$$\epsilon_0(h,b) = \int_{|\xi|>b} |\hat{h}(\xi)|\, d\xi \le Ce^{-\gamma b}$$

for certain $C > 0$, $\gamma > 0$. Equivalently, this means that $\|\mathcal{H}_b h\|_1$ decays exponentially as $b \to \infty$. By Hölder's inequality

$$\|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}} \le \|\hat{h}\|_\infty^{1/2} \|\mathcal{H}_b h\|_1^{1/2} \le Ce^{-\gamma b/2},$$

so for an essentially bandlimited function $h(x)$, $\|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}}$ also decays exponentially (as $b \to \infty$).

Corollary 4.2. *With the same notations as in the theorem, suppose that $b \to \infty$, $m \to \infty$, and either $m = \mathcal{O}(b)$ or $b = \mathcal{O}(m)$. Let $h(x)$ be essentially bandlimited. Then*

1. $\hat{g}_k = \mathcal{O}(|k|^{-\rho})$ *for some* $\rho > \frac{5}{2}$ *implies* $\|p - \mathcal{S}_{K,W}p\|_\infty = \mathcal{O}(m^{\frac{5}{2}-\rho}) \to 0$.
2. $\hat{g}_k = \mathcal{O}(|k|^{-\rho})$ *for some* $\rho > \frac{3}{2}$ *implies* $\|p - \mathcal{S}_{K,W}p\|_2 = \mathcal{O}(m^{\frac{3}{2}-\rho}) \to 0$.
3. $\hat{g}_k = \mathcal{O}(|k|^{-\rho})$ *for some* $\rho > 1$ *implies* $\|p - \mathcal{S}_{K,W}p\|_2 = \mathcal{O}(b^{\frac{1}{2}}m^{1-\rho})$.

Proof. Suppose $m = \mathcal{O}(b)$; then $m\epsilon_{-1}(h,b) = \mathcal{O}\big(b\epsilon_{-1}(h,b)\big) = \mathcal{O}(\epsilon_0(h,b))$. Similarly if $b = \mathcal{O}(m)$, then $m\epsilon_{-1}(h,b) = m\epsilon_{-1}(h,\mathcal{O}(m)) = \mathcal{O}(\epsilon_0(h,m))$. This shows that the terms involving $\epsilon_d(h,b)$ and $\|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}}$ in Theorem 4.1 decay exponentially, and we are left with the error terms related to g.

To estimate these, observe that for $\rho > 1$, $\hat{g}_k = \mathcal{O}(|k|^{-\rho})$ gives $R_k(g) = \mathcal{O}(|k|^{-\rho+\frac{1}{2}})$. Then

$$R_m[R_{(1-\vartheta')\cdot}(g)] = \mathcal{O}(m^{-\rho+1}),$$

which, using statement 3 in Theorem 4.1, gives the estimate 3. Similarly, if $\rho > 3/2$, then

$$\left(\sum_{\nu=1}^{\infty} \nu R_{(1-\vartheta')(\nu+m)}(g)^2\right)^{1/2} = \mathcal{O}\left(\left(\sum_{\nu=1}^{\infty} \nu(\nu+m)^{-2\rho+1}\right)^{1/2}\right) = \mathcal{O}(m^{-\rho+\frac{3}{2}}) \to 0,$$

which, using statement 2 in Theorem 4.1, provides estimate 2.

Finally, for $\rho > \frac{5}{2}$, $R_k(g) = \mathcal{O}(|k|^{-\rho+\frac{1}{2}})$ gives

$$\sum_{\nu=1}^{\infty} \nu R_{(1-\vartheta')(\nu+m)}(g) = \mathcal{O}\left(\sum_{\nu=1}^{\infty} \nu(\nu+m)^{-\rho+\frac{1}{2}}\right) = \mathcal{O}(m^{\frac{5}{2}-\rho}),$$

as claimed in statement 1. □

The estimates 1–3 do not include the case $\hat{g}_k = \mathcal{O}(|k|^{-1})$, as $R_{(1-\vartheta')(\nu+m)}(g)$ is then no longer well behaved. This may be overcome by considering different norms, as we proceed to do. For $1 < \alpha' < 2$ and $1/\alpha + 1/\alpha' = 1$, define a norm $|\cdot|_{\alpha'}$ on the physical plane by

$$|p|_{\alpha'} := \|\hat{p}\|_\alpha = \left(\sum_{k=-\infty}^{\infty} \int_{\mathbf{R}} |\hat{p}(\sigma,k)|^\alpha \, d\sigma\right)^{1/\alpha},$$

which is in accordance with the norms $\|\hat{p}\|_\alpha$ for $\alpha = 1, 2$ employed before. Notice that for $1 < \alpha' < 2$ these norms are less natural than the classical norms $\|\cdot\|_{\alpha'}$, but at least an estimate $|\cdot|_{\alpha'} \le \|\cdot\|_{\alpha'}$ holds (see [3, p. 177]), known as the Hausdorff–Young inequality. For $\alpha' > 2$, the Hausdorff–Young inequality is no longer true, and usage of the norms $|\cdot|_{\alpha'}$ would then appear rather airy.

Corollary 4.3. *With the same hypothesis as in Corollary* 4.2*, let* $\hat{g}_k = \mathcal{O}(|k|^{-1})$*, and suppose* $h(x)$ *is essentially bandlimited.*

4. *If* $1 < \alpha' < \frac{4}{3}$*, then* $|p - \mathcal{S}_{K,W}p|_{\alpha'} = \mathcal{O}(m^{\frac{3\alpha'-4}{4\alpha'-4}}) \to 0$.
5. *If* $1 < \alpha' < 2$*, then* $|p - \mathcal{S}_{K,W}p|_{\alpha'} = \mathcal{O}(b^{\frac{\alpha'-1}{\alpha'}} m^{\frac{1}{2}-\frac{1}{\alpha'}})$.

Proof. We have to go through the proof of Theorem 4.1 with the norm $\|\hat{p} - \widehat{\big(\mathcal{S}_{K,W}p\big)}\|_\alpha$ as in (4.5), (4.6), but with different $\alpha > 2$. We estimate on the domains Γ_1, Γ_2. Now according to (4.12),

$$\|\chi_{\Gamma_2}\hat{p}\|_\alpha \le \|\hat{p}\|_\infty^{1-2/\alpha} \|\chi_{\Gamma_2}\hat{p}\|_2^{2/\alpha} \le C\|\mathcal{H}_b h\|_{W^{2,-\frac{1}{2}}}^{2/\alpha},$$

which decays exponentially, since h is essentially bandlimited. So we are left with the error estimate on Γ_1 involving the error terms for g.

Proceeding as in part 8 of the proof of Theorem 4.1, we obtain

$$\|\chi_{\Sigma_1}\hat{p}\|_\alpha^\alpha \leq C \sum_{|k|\geq m} (|k|-m)\bigg(R_{(1-\vartheta')|k|}(g)+\|g\|_2 e^{-\gamma|k|}\bigg)^\alpha = \mathcal{O}\bigg(\sum_{\nu=1}^\infty \nu R_{(1-\vartheta')(\nu+m)}(g)^\alpha\bigg),$$

the exponentially decaying term being negligible. Now obviously $\hat{g}_k = \mathcal{O}(|k|^{-1})$ implies $R_{(1-\vartheta')|k|}(g) = \mathcal{O}(|k|^{-1/2})$, so for $\alpha > 4$,

$$\sum_{\nu=1}^\infty \nu R_{(1-\vartheta')(\nu+m)}(g)^\alpha = \mathcal{O}\bigg(\sum_{\nu=1}^\infty \nu(\nu+m)^{-\alpha/2}\bigg) = \mathcal{O}(m^{2-\alpha/2}) \to 0,$$

giving $|p - \mathcal{S}_{K,W}p|_{\alpha'} = \mathcal{O}(m^{1-\alpha/4})$, which is estimate 4. Finally, if only $\alpha > 2$, we still have $\sum_\nu R_{(1-\vartheta')(\nu+m)}(g)^\alpha = \mathcal{O}(m^{1-\frac{\alpha}{2}})$, which readily gives estimate 5. This completes the proof. □

The principal message of Theorem 4.1 and its corollaries is that the aliasing error associated with a choice of the bowtie region (4.3) may be attributed to two different sources. The errors on the regions Σ_2, Σ_3 (resp., Γ_2) decay exponentially (as $b \to \infty$) if $h(x)$ is essentially bandlimited. On the other hand, the error contribution from Σ_1, and correspondingly, from Γ_1, entirely depends on g and no longer relates to the spatial bandwidth b. This error contribution decays as $m \to \infty$, but in general much slower than the other error terms. In practice, this may require choosing a rather large m, which may render an appropriate sampling difficult (cf. Figure 4). In detail, we have the following observations.

Remarks. (1) If g is of class $\mathcal{C}_{\text{per}}^\infty$ or even analytic, the error from region Σ_1, and hence the overall aliasing error, decays rapidly as the support region K grows. This is of course the case when the source is static, so we reproduce Natterer's estimates in [20, Thm. III.3.1]. Similarly, if $g(t)$ presents a full dynamic profile, starting with 0 activity, reaching its peak after uptake, and decaying back to 0 after washout, we may realistically assume that $g((T/2\pi)\phi) \in \mathcal{C}_{\text{per}}^\infty$, which again gives a fast decay as $m \to \infty$.

(2) In many practical cases, however, $g((T/2\pi)\phi)$ is not even of class $\mathcal{C}_{\text{per}}$. For instance if only a washout profile is scanned, we usually find $g(t)$ decaying like an exponential or a sum of exponentials, so $g((T/2\pi)\phi)$ is piecewise analytic but discontinuous. Here Theorem 4.1 and Corollary 4.2 are not applicable, since the Fourier coefficients of $g(t) = e^{-\lambda t}$ are $\hat{g}_k = \mathcal{O}(|k|^{-1})$. In this case, we have to retreat to the estimate 4 from Corollary 4.3, which is not entirely satisfactory, as it involves a norm $|\cdot|_{\alpha'}$ with $1 < \alpha' < 2$. One may very well argue that failure of 2-norm convergence indicates a problem in practice, and some of our experiments seem to emphasize this (cf. Figure 4).

(3) Notice that $\hat{g}_k = \mathcal{O}(|k|^{-2})$ if, according to the terminology of [3], g satisfies a *generalized Lipschitz condition of order* 2, that is, if

$$g(\phi+h) + g(\phi-h) - 2g(\phi) = \mathcal{O}(h^{-2}) \qquad \text{as } h \to 0, \tag{4.19}$$

uniformly in $\phi \in [0, 2\pi]$.

(4) The first 2-D Fourier analysis of the unattenuated Radon transform was presented in [22]. These authors calculate the spectrum of a point source $f(x) = \delta(x-a)$. Following their idea, one might consider a *dynamic point source*

$$f(t,x) = g(t)\,\delta(x-a) \tag{4.20}$$

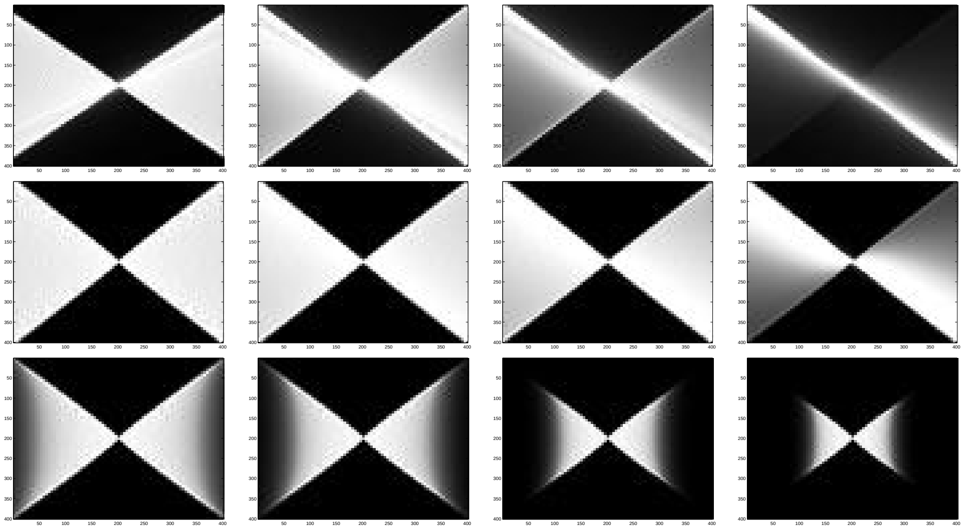


FIG. 3. *The first two lines show the effect of the dynamic* $g(t) = e^{-\lambda(2\pi/T)t}$ *on a point source located at* $a = (0.56, 0.8285) \in D$*. The first line displays the cases* $\lambda = 0.1, 0.4, 0.6,$ *and* 1.1 *(left to right). For a fast decay, the spectrum tends to emphasize a diagonal with slope related to the position* a *of the source. The second line shows the energy spectra for the same dynamics, but after doubling the data. The third line shows the effect of the spatial bandwidth* b*. This may be simulated by considering sources of the form* $h(x) = \phi_{a,\sigma}(x)$ *with* $g \equiv 1$*, where* $\phi_{a,\sigma}$ *denotes the* 2*-D Gaussian with mean* $a = (0.56, 0.8285) \in D$ *and covariance matrix* $\sigma^2 I_2$ *for different* $\sigma = 0.008, 0.01, 0.015,$ *and* 0.02 *(left to right). For* σ *not exceedingly large,* $\phi_{a,\sigma}$ *may be considered as compactly supported. Notice that* $R\phi_{a,\sigma}(s,\omega) = \phi_{a\cdot\omega^\perp}(s)$ *is a* 1*-D Gaussian, whose spectrum may be calculated analytically, cf.* [15].

located at $a \in D$ and emitting with dynamic profile $g(t)$. In fact, the energy spectrum of (4.20), while obviously not bounded in σ-direction, still decays on the region Σ_1 (as $m \to \infty$). This leads to a support region of infinite bowtie shape (see Figure 3). The point of view adopted by considering sources (4.20) is useful since it directly relates the thickness of the bowtie at $\sigma = 0$ to the dynamic profile $g(t)$.

(5) The analysis presented in Theorem 4.1 breaks down at an early stage when R is replaced by the attenuated Radon transform. Even the Fourier slice theorem is no longer available, nor has it an equally useful alter ego. This seems to limit the analysis to numerical experiments, which is of course not entirely satisfactory. Fortunately, adopting the point of view expounded in remark 4, we may interpret attenuation as a particular type of dynamics. In fact, consider a dynamic point source (4.20) located at a and attenuated through $\mu(x)$. Define the function $g_a(\phi)$ by

$$g_a(\phi) = \exp\Big\{ -\int_0^\infty \mu(a + \tau\omega)\, d\tau \Big\}, \qquad \omega = (\cos\phi, \sin\phi).$$

The attenuated Radon transform of $\delta(\cdot - a)$ is

$$\begin{aligned} R[\mu, \delta(\cdot - a)](s,\phi) &= \int_{-\infty}^\infty \delta(s\omega^\perp + \tau\omega - a)\, \exp\Big\{ -\int_\tau^\infty \mu(s\omega^\perp + \tau'\omega)\, d\tau' \Big\}\, d\tau \\ &= \exp\Big\{ -\int_{a\cdot\omega}^\infty \mu(s\omega^\perp + \tau'\omega)\, d\tau' \Big\}\, \delta(s - a\cdot\omega^\perp) \\ &= g_a(\phi)\, \delta(s - a\cdot\omega^\perp) \;=\; g_a(\phi)\, R\,\delta(\cdot - a)(s,\phi). \end{aligned}$$

Therefore, for a single camera head, according to (3.1), the sinogram of (4.20) is

$$p(s,\phi) = g((T/2\pi)\phi)\, g_a(\phi)\, R\,\delta(\cdot - a)(s,\phi). \tag{4.21}$$

The interpretation of (4.21) is that on a static point source, attenuation acts like a dynamic, while for a dynamic point source, it modulates the existing dynamics. The important point, however, is that as long as 360 degrees are scanned, $g_a(\phi)$ is smooth (as soon as $\mu(x)$ is). Therefore, one may argue that modulating the existing dynamics will not seriously slow down the convergence of the Fourier series, and attenuation will not qualitatively alter the shape of the infinite bowtie support region of the dynamic point source with profile $g(t)$. This seems to be corroborated by numerical experiments.

(6) Notice that a result similar to Theorem 4.1 and Corollary 4.2 may be obtained on a 180-degree tour. The estimates involving Bessel functions have to be modified, but the coefficients replacing $J_n(\theta n)$ still decay exponentially. What makes a 180-degree tour seem more delicate is the more serious effect of attenuation. Namely, $g_a(\phi)$, defined on $[0,\pi]$, and continued periodically outside, will now just like $g((T/\pi)\phi)$ have a discontinuity at $\phi = 0$, adding to the effect of the discontinuity of $g((T/\pi)\phi)$ at $\phi = 0$. Doubling the data in the way shown in the next section will partially remedy this (see Figure 4).

5. Experiments.

While the results in the previous section serve to theoretically justify the choice of a frequency window of bowtie shape, K, they do not readily indicate how to calculate K (or rather, m and b) in practice. To do this, we have to provide a practical guideline. Treating the error contributions from Σ_1 and Σ_2, Σ_3 separately, we propose the following approach.

For a dynamic profile $g(t)$ having $\hat{g}_k = \mathcal{O}(|k|^{-\rho})$ for some $\rho > 1$, and for a point source $\delta(\cdot - a)$ located on the unit disk D, consider the spectrum $\hat{p}$ of the sinogram p of $f(t,x) = g(t)\,\delta(x-a)$. For every frequency σ choose indices $\underline{m}(\sigma)$ and $\overline{m}(\sigma)$ such that

$$\sum_{\nu=\underline{m}(\sigma)}^{\overline{m}(\sigma)} |\hat{p}_\nu(\sigma)|^2 \;\geq\; .98^{\frac{1}{2}} \cdot \sum_{\nu=-\infty}^{\infty} |\hat{p}_\nu(\sigma)|^2, \tag{5.1}$$

uniformly over $a \in D$, which is to say that on each line $\sigma = \text{const}$, $[\underline{m}(\sigma), \overline{m}(\sigma)]$ captures 98.99% of the energy of $\hat{p}(\sigma,\cdot)$ (notice $.9899 = .98^{\frac{1}{2}}$). This procedure will, if successful for a given dynamic $g(t)$, provide an infinite region which essentially captures the energy of the spectrum $\hat{p}$ of any source of the form $f(t,x) = g(t)\,h(x)$, $h(x)$ supported on D, but not necessarily bandlimited. It is hoped that the *same* region will then emerge for a large variety of dynamic profiles $g(t)$.

As it turns out, this program is indeed realizable. Numerical experiments indicate that the desired infinite support region is an infinite bowtie as displayed in Figure 3, with a symmetry $\underline{m} = -\overline{m}$ apparent. In (4.3), the choice $\vartheta \approx 1$ seems justified, and for a large variety of profiles $g(t)$, the delimiters $\underline{m}$, $\overline{m}$ are then of the form

$$\overline{m}(\sigma) = |\sigma| + m, \quad \underline{m} = -\overline{m},$$

where $m = \overline{m}(0) > 0$ may be calculated explicitly through (5.1).

In a second step we now have to truncate the infinite bowtie at $\sigma = \pm b$ in order to define a bounded region K. This may be done by applying the same argument again, i.e., by choosing b to satisfy

$$\sum_{\nu=\underline{m}(\sigma)}^{\overline{m}(\sigma)} \int_{|\sigma|\le b} |\hat{p}_\nu(\sigma)|^2 \, d\sigma \;\ge\; .98^{\frac{1}{2}} \cdot \sum_{\nu=\underline{m}(\sigma)}^{\overline{m}(\sigma)} \int_{\mathbf{R}} |\hat{p}_\nu(\sigma)|^2 \, d\sigma \tag{5.2}$$

uniformly over $a \in D$. Clearly this step may require a discretization, conveniently done by FFT 2. The combined procedure (5.1), (5.2) specifies a region K carrying 98% of the energy of the spectrum $\hat{p}$.

A second and easier way to specify a bowtie (4.3) is to discretize the spectrum $\hat{p}$ into a frame of size $S \times S$, say, and then find a tolerance $\epsilon > 0$ to the effect that within the chosen frame, $\hat{p}\chi_{\{|\cdot|\ge\epsilon\}}$ carries 98% of the energy of $\hat{p}$. Both procedures turn out to be in good agreement, which reinforces the choice (4.3).

Let us now consider an application with experimental data, exhibiting the typical problem with a discontinuous time profile. The study shown in Figure 4 uses a phantom built at Vancouver General Hospital [4] and was performed with a Siemens Multispect-3 (MS3) triple head camera with a low energy ultra high resolution (LEUHR) collimator. Only data from one of the camera heads were used to simulate the case of a single head camera.

The phantom, a 17-ml container shown in Figure 4(a), is connected to a supply and a drain and equipped with a mixing propeller to guarantee a homogeneous flow. The container was initially filled with Tc-99 m of approximately 40 MBq radio activity. The activity was diluted and washed out through uniform water flow, producing approximately a single exponential decay with estimated half-life of 3 minutes. The plot of the total activities of the 64 views of a slice selected at the horizontal pixel position 38 is shown in Figure 4(b). The sinogram of the selected slice is shown in (c), indicating that 180 degrees have been scanned with 64 stops and a camera cross section divided into 64 bins. The time for the total scan was $T = 10$ minutes.

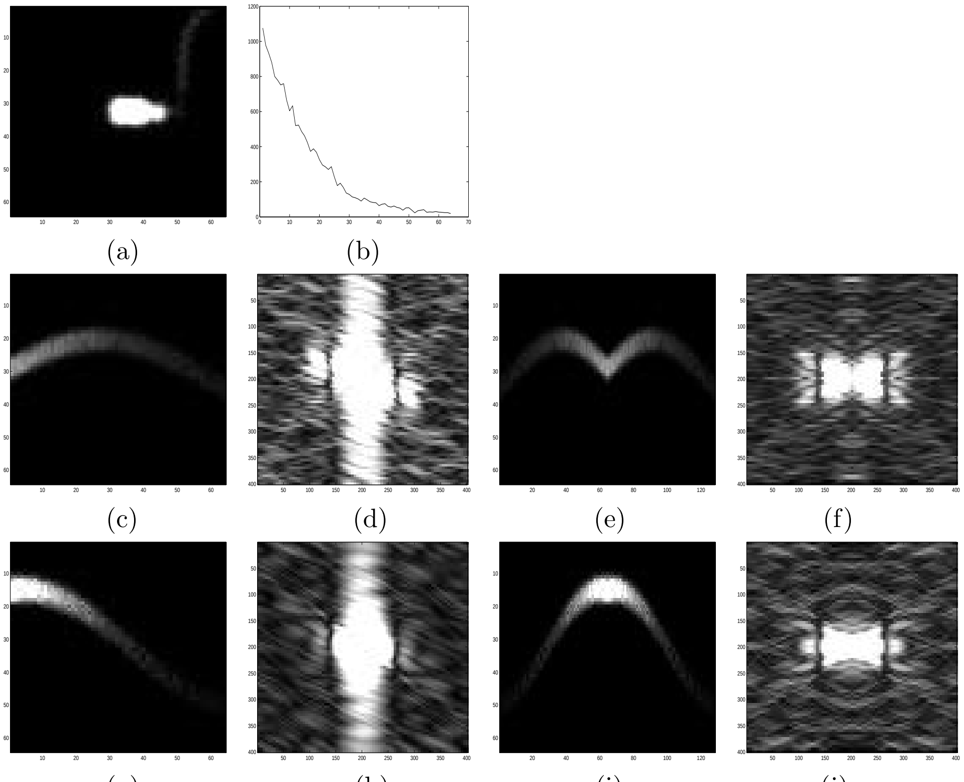
Figure 4(d) shows the energy spectrum of Figure 4(c), obtained via zero filling into a 400×400 frame, applying the 2-D FFT, taking absolute values, and rearranging the image so that high frequencies are at the edges.

The energy spectrum Figure 4(d), expected to resemble a bowtie shape, is blotted by a high energy band in vertical direction. According to theory, this high energy band should not exist here—unless some of the hypothesis on which the results in section 4 are based turned out to be violated. As a list of possible explanations we offer the following.

(1) The sinogram, being blurred by attenuation and scatter, may contain noise components not modeled in (3.2), whose spectra contribute to the vertical band.

(2) The bowtie region (4.3) was obtained under the hypothesis that the object is contained in the unit disk. While this is the case for the selected slice, we have to remember that the radiating object is 3-D, and neighboring slices contribute to the data through scatter and collimator blurring. Some of these recorded events may be mistaken as coming from outside the unit circle.

(3) The problem evoked before: the recorded data present a washout with approximately single exponential decay, see Figure 4(b), causing a discontinuity of $g((T/\pi)\phi)$ at $\phi = 0$. The high energy band visible in Figure 4(d) may indicate the failure of convergence of the Fourier series at $\phi = 0$, or rather, that a very large $m = \overline{m}(0)$ is required in (5.1).



(a) (b)

(c) (d) (e) (f)

(g) (h) (i) (j)

FIG. 4. *Experimental data from Vancouver General Hospital, obtained with a Siemens Multispect-*3 *triple head camera.*

(4) Since a 180-degree sector has been scanned, the sinogram, along with the discontinuity in time, has a singularity in the spatial variable, visible in Figure 4(e) as a kink, which may as well be responsible for the phenomenon.

In order to decide which of these items is likely to cause the phenomenon of Figure 4(c), we *double* the data by flipping the sinogram with respect to the axis $\phi = 0$, including the reverse data among a new symmetric sinogram of size 128×64, displayed in Figure 4(e). The doubled sinogram now has an increase of activity on $[-\pi, 0]$, followed by the original period of decay on $[0, \pi]$.

The effect of the doubling procedure, while theoretically improving the signal to be of class $\mathcal{C}_{\mathrm{per}}(-\pi, \pi)$, is dramatic in the case of our experiment. The energy spectrum Figure 4(f) of the doubled sinogram no longer exhibits the erratic vertical energy band and quite reasonably displays the bowtie form predicted by theory. As the decay profile of the bottle may very well be approximated by an exponential $e^{-\lambda\phi}$, the doubling procedure may even theoretically be justified. While the Fourier coefficients of $e^{-\lambda\phi}$ are $\mathcal{O}(|k|^{-1})$, the doubled signal $e^{-\lambda|\phi|}$ on $[-\pi, \pi]$ has Fourier coefficients $c_k = (1 + (-1)^{k+1} e^{-\lambda\pi}) \frac{2\lambda}{\lambda^2 + k^2} = \mathcal{O}(|k|^{-2})$. In any case, we strongly recommend the doubling procedure, particularly if the dynamic is relatively fast.

In order to indicate that the phenomenon in Figure 4(d) is not due to any of the noise effects evoked in items (1) and (2) of our list, one may create an artificial 2-D object, resembling the true slice of the bottle, with activity distribution a properly scaled 2-D Gaussian. For the dynamics one would substitute a single exponential decay found by inspecting the cumulative activity plot Figure 4(b). The result, not displayed here, shows that the Gibbs phenomenon is still apparent, reinforcing our explanation (3).

Finally, to discriminate item (3) from the possible explanation (4), we have scanned the bottle from a different position (Figure 4(a)), the sinogram of the corresponding slice shown in Figure 4(g), and selected to the effect that in the doubled sinogram in Figure 4(i) the spatial singularity is removed. The energy spectra in Figure 4(h), belonging to in Figure 4(g), and in Figure 4(j), belonging to Figure 4(i), show that the result is qualitatively the same, indicating that the phenomenon (4) is less serious than (3).

6. Resolution. We present the promised guideline on how to acquire data with a rotating SPECT camera. Consider exemplarily the case of a single camera head rotating over a 180-degree tour. Doubling the data will then give a 360-degree sinogram. Suppose the dynamic source $f(t,x)$ is of the form $g_1(t)\,h_1(x)+\cdots+g_r(t)\,h_r(x)$, with the $h_i(x)$ supported on the unit disk. Assume that the unit circle is completely visible from each camera position, which means that a camera cross section has length 2. Assume that the cross section is divided into 64 bins, giving $\Delta s = 1/32$. Since the Nyquist rate is $\Delta s = \pi/b$, the best possible bandwidth is $b = 32\pi$, a fact we may not easily debate if the resolution of the camera has to be considered a fixed technical parameter.

According to (4.4), the sampling parameters in the frequency plane are $\Delta\sigma = b$, $\Delta k = [b]+2m$, where we have chosen $\vartheta \approx 1$, as validated by the numerical experiments in section 4. Using (4.4), this gives

$$\Delta s = \frac{\pi}{b}, \qquad \Delta\phi = \frac{\pi}{[b]+2m}. \tag{6.1}$$

Let us consider the case where a washout (with decreasing activity) is scanned. As it comes out, tracer dynamics are often described by a *compartmental model* (cf. [11]), and accordingly the dynamic source is represented as a sum of exponentials

$$f(t,x) = h_1(x)\,e^{-\lambda_1(\pi/T)t} + \cdots + h_r(x)\,e^{-\lambda_r(\pi/T)t}$$

with $\lambda_i \geq 0$ (decay to 0 at infinity) and where the $h_i(x)$ are supported on the unit disk D. In practice we may usually exhibit $\lambda_i \leq \lambda$, the fastest dynamic to be expected. Then the bowtie region may be estimated by considering a source of the form $f(t,x) = h(x)\,g(t) = h(x)\,e^{-\lambda(\pi/T)t}$.

As $g((T/\pi)\phi) = e^{-\lambda\phi}$, doubling the data as suggested in our approach gives the dynamic profile $e^{-\lambda|\phi|}$ over $-\pi \leq \phi \leq \pi$. Estimating the thickness $m = m(\lambda)$ of the bowtie as a function of λ, based on (5.1), yields the approximate linear relationship

$$m(\lambda) \approx \frac{3}{8}\,\lambda + 1, \tag{6.2}$$

which we exploit a little further by considering a realistic situation comparable to the one in our experiment.

Suppose that the total acquisition time of the scan is $T = 10$ minutes, while the shortest expected half-life is of the order of 2 minutes. Then $t_{\frac{1}{2}} = T\log 2/\pi\lambda = .2T$, giving $\lambda = \log 2/.2\pi \approx 1.1$. Hence $g((T/\pi)\phi) = \exp\{(-\log 2/.2\pi)|\phi|\} = \exp\{-1.1|\phi|\}$, which through (6.2) suggests $m \approx 1.4$. In view of (6.1), this gives $\Delta\phi \approx .06 = 3.44$

degrees as satisfactory for practical purposes. The interpretation is that an appropriate sampling of the doubled signal over 360 degrees requires approximately 105 views, that is, 53 views over 180-degrees. This is not in complete agreement with the actual policies (cf. [16]), where we often prefer to take 64 views on a 180-degree tour. As our scenario ignores the noise contributions, 53 views may in practice be barely sufficient, and we consider 64 views as a practical guideline on 180 degrees.

Remark. Formula (6.1) may be interpreted as an *uncertainty principle*. Assume that a minimum $\Delta\phi$ has been specified by the user to guarantee a sufficient number of recorded counts per camera position. Then we may consider $[b] + 2m$ as fixed. So within certain limits, we may either increase m and capture faster dynamics, paying eventually by a loss in spatial resolution (by decreasing b), or we may conversely choose a better spatial resolution by increasing b, bearing the risk that some of the faster dynamics may not be adequately represented.

7. Filtering. In this last section we discuss a policy for the 2-D filtering of the sinogram data. Notice that filtering of the projection data is currently done in one dimension, that is, every projection is filtered separately. In the static case, this does not cause any particular difficulty, as the same filter may be used for all projections. To that effect, various filters have been around for years, and their application is well understood.

The situation is a little more complicated in dSPECT, as the overall activity changes from view to view. 1-D filtering may now require adapting an individual filter to each projection, and it may then seem more attractive to do a 2-D filtering, based on the insights of section 4. In particular, a 2-D filter, if based on the 2-D Fourier transform, may use the bowtie shape of the spectrum of the Radon transform, and may therefore incorporate information not easily assessed through a 1-D procedure. We therefore propose the following frequency domain based 2-D filtering procedure, which incorporates the theoretical results obtained in previous sections.

To render the situation even more interesting, we modify the experiment from section 5 by scanning four bottles of the type shown in Figure 4(a). We arrange a washout through continuous water flow of different half-lives between 2 and 6 minutes. Starting out with the 64×64 sinogram (Figure 5, top left), we double data (top row right) as done previously, and include them into a frame of zeros of considerably larger size $L \times L$ (zero filling), where usually $L = 200$ or $L = 400$. The 2-D FFT is applied to the enlarged signal. Filtering is now performed in the frequency domain by multiplying the $L \times L$ spectrum with a 2-D window function:

$$w_{m,b}(x, y) = \phi(x/b)\,\phi(y/(m + x))$$

with $-L/2 \le x, y \le L/2$ integer and parameters $m, b \le L/2$. The window function satisfies $\phi(0) = 1$, $\phi(t) = 0$ for $|t| \ge 1$ and could be any of the standard 1-D lag windows. Figure 5 (second row) displays several filters with the choice $L = 200$, b ranging from 100 to 70, and m ranging from 80 to 20.

Rows 3–6 show the effect of the 2-D filtering of these window functions. The left-hand picture shows three projections (number 4, number 19, and number 61). The dotted line shows the original data, the continuous line shows the filtered curves. The right-hand diagram shows the smoothing effect of the filters on the sum plot. The latter indicates the success of the doubling, as the same filtering applied to the simple sinogram would exhibit a Gibbs phenomenon at $\phi = 0$.

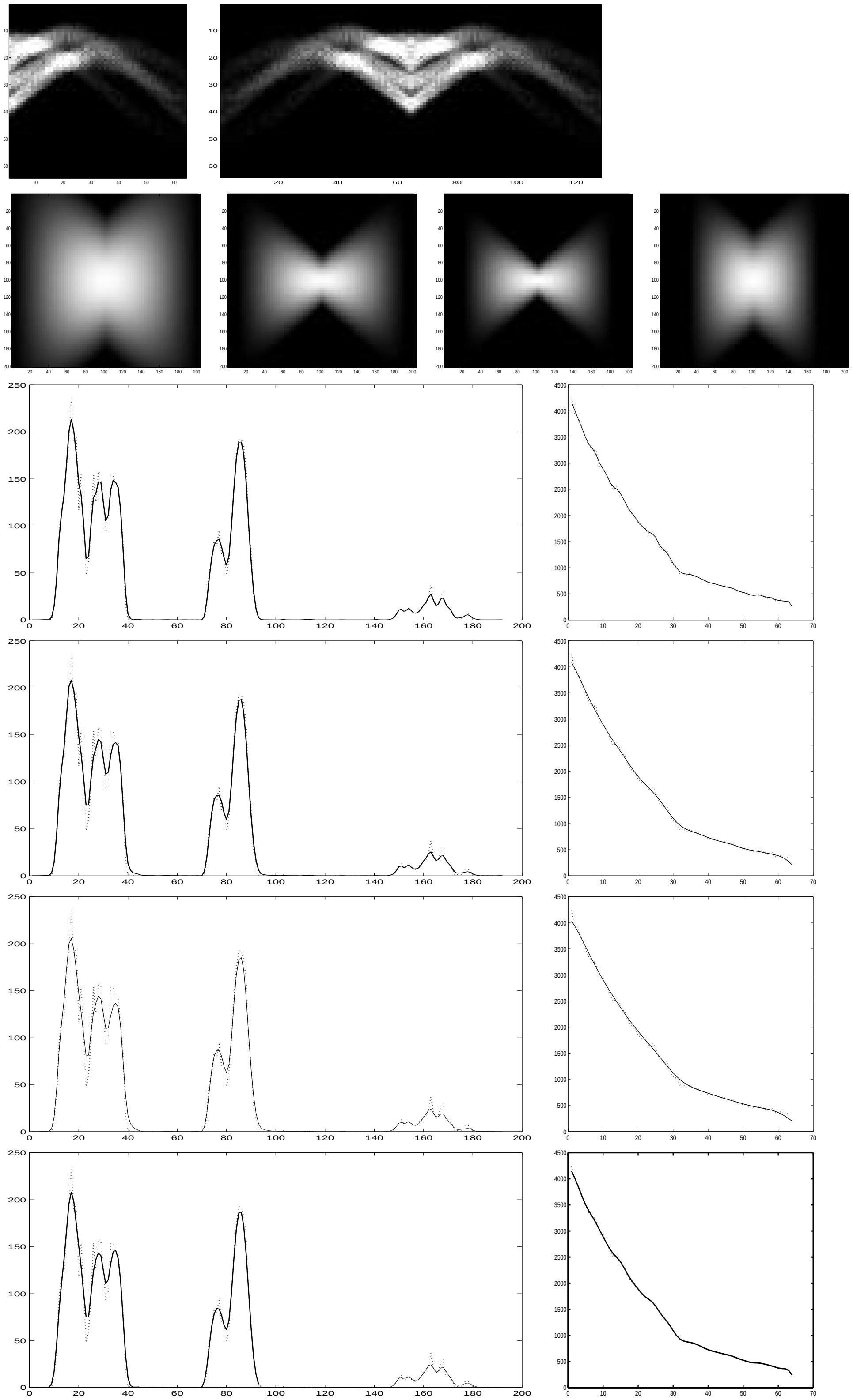


Fig. 5. *2-D filtering of experimental data.*

REFERENCES

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] H.H. Bauschke, D. Noll, A. Celler, and J.M. Borwein, *An EM-algorithm for Dynamic SPEC Tomography*, IEEE Trans. Med. Imag., 18 (1999), pp. 252–261.

[3] P.L. Butzer and R.J. Nessel, *Fourier Analysis and Approximation*, Vol. 1, Birkhäuser-Verlag, Basel, 1971.

[4] A. Celler, T. Farncombe, R. Harrop, and D. Lyster, *Dynamic heart-in-thorax phantom for functional SPECT*, IEEE Trans. Nuclear Sciences, 44 (1997), pp. 1600–1605.

[5] A. Celler, T. Farncombe, R. Harrop, D. Noll, and J. Maeght, *Dynamic SPECT imaging using single camera rotation (dSPECT)*, IEEE Trans. Nuclear Sciences, 46 (1999), pp. 1055–1061.

[6] A. Celler, A. Sitek, and R. Harrop, *Reconstruction of multiple line source attenuation maps*, in IEEE Nuclear Science Symposium Conference Record, IEEE, New York, 1996, pp. 1420–1424.

[7] A. Celler, A. Sitek, E. Stoub, D. Lyster, C. Dykstra, D. Worsley, and A. Fung, *Development of a multiple line source attenuation array for SPECT transmission scans*, J. Nuclear Medicine, 38 (1997), p. 215.

[8] R. Dautrey and J.-L. Lions, *Analyse mathématique et calcul numérique*, Tome 9, Masson, Paris, 1984.

[9] V. Dicken, *Simultaneous activity and attenuation reconstruction in emission tomography*, Inverse Problems, 15 (1999), pp. 931–960.

[10] G.T. Gullberg, R.H. Huesman, S.G. Ross, E.V.R. Di Bella, G.L. Zeng, B.W. Reutter, P.E. Christian, and S.A. Foresti, *Dynamic cardiac single photon emission computed tomography*, in Nuclear Cardiology: State of the Art and Future Directions, Mosby-Year Book, Philadelphia, in press.

[11] A. Gjedde, *Compartmental Analysis*, in Principles of Nuclear Medicine, 2nd ed., H.N. Wagner, Jr., Z. Szabo, J.W. Buchanan, Saunders, Philadelphia, PA, 1995, pp. 451–461.

[12] C.M. Kao, J.T. Yap, J. Mukherjee, M. Cooper, and C.T. Chen, *An image reconstruction method for dynamic PET*, IEEE NSS/MIC Conference Record, IEEE, New York, 1995.

[13] H. Kruse, *Resolution of reconstruction methods in computerized tomography*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 447–474.

[14] M.A. Limber, M.N. Limber, A. Celler, J.S. Barney, and J.M. Borwein, *Direct reconstruction of functional parameters for dynamic SPECT*, IEEE Trans. Nuclear Sciences, 42 (1995), pp. 1249–1256.

[15] A. Kuruc, *Efficient estimation of linear functionals in emission tomography*, SIAM J. Appl. Math., 57 (1997), pp. 426–452.

[16] J. Maeght, D. Noll, A. Celler, and T. Farncombe, *Methods for dynamic SPECT*, Rapport Interne 99-26, Laboratoire Mathématiques pour l'Industrie et la Physique, 1999. Also available at http://mip.ups.-tlse.fr/publi.

[17] K. Nakajima, J. Taki, H. Bunko, M. Matsudaira, A. Muramori, I. Matsunari, K. Hisada, and T. Ichihara, *Dynamic acquisition with a three-headed SPECT system: Application of Technetium 99m-SQ30217 myocardial imaging*, J. Nuclear Medicine, 32 (1991), pp. 1273–1277.

[18] F. Natterer, *Computerized tomography with unknown sources*, SIAM J. Appl. Math., 43 (1983), pp. 1201–1212.

[19] F. Natterer, *Sampling in fan beam tomography*, SIAM J. Appl. Math., 53 (1993), pp. 358–380.

[20] F. Natterer, *The Mathematics of Computerized Tomography*, Teubner-Verlag, Stuttgart, 1986.

[21] F. Natterer, *Determination of tissue attenuation in emission tomography of optically dense media*, Inverse Problems, 9 (1993), pp. 731–736.

[22] P.A. Rattey and A.G. Lindgren, *Sampling the 2-D Radon transform*, IEEE Trans. Acoust. Speech Signal Process., 29 (1981), pp. 994–1002.

[23] G.N. Watson, *A Treatise on Bessel Functions*, Cambridge University Press, Cambridge, UK, 1952.

[24] A. Welch, R. Clark, F. Natterer, and G.T. Gullberg, *Toward accurate attenuation correction in SPECT without transmission measurements*, IEEE Trans. Med. Imag., 16 (1997), pp. 532–541.

[25] R.G. Wells, A. Celler, and R. Harrop, *Analytic calculation of photon distribution in SPECT projections*, IEEE Trans. Nuclear Sciences, 45 (1998), pp. 3202–3214.

SIAM J. Math. Anal.
Vol. 31, No. 5, pp. 1121–1143

SECOND ORDER SINGULAR PERTURBATION MODELS FOR PHASE TRANSITIONS[*]

IRENE FONSECA[†] AND CARLO MANTEGAZZA[‡]

Abstract. Singular perturbation models involving a penalization of the first order derivatives have provided a new insight into the role played by surface energies in the study of phase transitions problems. It is known that if $W : \mathbb{R}^d \to [0, +\infty)$ grows at least linearly at infinity and it has exactly two potential wells of level zero at $a, b \in \mathbb{R}^d$, then the $\Gamma(L^1)$-limit of the family of functionals

$$\mathcal{F}_\varepsilon(u) := \begin{cases} \int_\Omega \left(\frac{W(u)}{\varepsilon} + \varepsilon|\nabla u|^2\right) dx & \text{if } u \in W^{1,2}(\Omega; \mathbb{R}^d), \\ \\ +\infty & \text{if } u \in L^1(\Omega; \mathbb{R}^d) \setminus W^{1,2}(\Omega; \mathbb{R}^d), \end{cases}$$

where Ω is a bounded, open set in $\mathbb{R}^N$, is given by

$$\mathcal{F}(u) := \begin{cases} \mathbf{m}\,\mathrm{Per}_\Omega(\{u = a\}) & \text{if } u \in BV(\Omega; \{a, b\}), \\ +\infty & \text{otherwise,} \end{cases}$$

for a suitable constant $\mathbf{m}$ depending on the energy density W. In this paper, and motivated by the study of phase transitions for nonlinear elastic materials, the $\Gamma(L^1)$-limit is obtained in the case where in $\mathcal{F}_\varepsilon(u)$ the penalization term $\varepsilon|\nabla u|^2$ is replaced by $\varepsilon^3|\nabla^2 u|^2$, for $u \in W^{2,2}(\Omega; \mathbb{R}^d)$. The resulting functional is of the same form as $\mathcal{F}(u)$ above.

Key words. Γ-limit, interpolation inequalities, Young measures

AMS subject classifications. 49J45, 49Q20

PII. S0036141099356830

1. Introduction.

In this paper we show that the $\Gamma(L^1)$-limit of the family of singular perturbations

$$\mathcal{F}_\varepsilon(u) := \begin{cases} \int_\Omega \left(\frac{W(u)}{\varepsilon} + \varepsilon^3|\nabla^2 u|^2\right) dx & \text{if } u \in W^{2,2}(\Omega; \mathbb{R}^d), \\ \\ +\infty & \text{if } u \in L^1(\Omega; \mathbb{R}^d) \setminus W^{2,2}(\Omega; \mathbb{R}^d), \end{cases}$$

where $W : \mathbb{R}^d \to [0, +\infty)$ grows at least linearly at infinity and has exactly two potential wells of zero level at $a, b \in \mathbb{R}^d$, is given by

$$\mathcal{F}(u) := \begin{cases} \mathbf{m}\,\mathrm{Per}_\Omega(\{u = a\}) & \text{if } u \in BV(\Omega; \{a, b\}), \\ +\infty & \text{otherwise,} \end{cases}$$

with

$$\mathbf{m} := \min\left\{ \int_{\mathbb{R}} (W(f) + |f''|^2\, dt : f \in W^{2,2}_{\mathrm{loc}}(\mathbb{R}; \mathbb{R}^d), \lim_{t\to+\infty} f(t) = b, \lim_{t\to-\infty} f(t) = a \right\}.$$

[*] Received by the editors June 7, 1999; accepted for publication (in revised form) December 8, 1999; published electronically May 2, 2000.
http://www.siam.org/journals/sima/31-5/35683.html

[†] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (fonseca@andrew.cmu.edu). The research of this author was partially supported by the National Science Foundation under grants DMS-9500531 and DMS-9731957 and through the Center for Nonlinear Analysis (CNA).

[‡] Scuola Normale Superiore, Pisa, 56126, Italy (mantegaz@sns.it).

Singular perturbations of nonconvex, multiple-well variational problems may be found in gradient strain theories in plasticity, ferromagnetics, and other areas of materials science and engineering. In particular, within the context of phase transitions of nonlinear elastic materials, let $W : \mathbb{R}^{d\times N} \to [0, +\infty)$ be the stored energy density of a material with reference configuration an open, bounded set $\Omega \subset \mathbb{R}^N$, and which may undergo a phase transformation. This material instability may be due, in part, to the multiple-well profile of W. For simplicity, assume that

$$W(\xi) = 0 \text{ if and only if } \xi \in \{A, B\},$$

where $\operatorname{rank}(A - B) = 1$. Let us assume further that equilibria for fixed phase volume fraction are determined by minimum energy; physically, this model is oversimplified since it is incompatible with the frame indifference requirement; it does not take into account material symmetries, evolution is neglected, and there is no heat diffusion. Then we are led to (see [20, 22])

$$\min\left\{\int_\Omega W(\nabla u)\,dx : u \in W^{1,1}(\Omega; \mathbb{R}^d), \int_\Omega \nabla u\,dx = \mathcal{L}^N(\Omega)(\theta A + (1-\theta)B)\right\},$$

where $\theta \in (0, 1)$ is a fixed volume fraction, and $\mathcal{L}^N$ stands for the N-dimensional Lebesgue measure in $\mathbb{R}^N$. Due to the rank-one compatibility between A and B, there are infinitely many laminates with strain gradients alternating between A and B which will minimize the total bulk energy. As in the Cahn–Hilliard model for liquid-liquid phase transformations with underlying variational formulation

$$\min\left\{\int_\Omega W(u)\,dx : u \in L^1(\Omega; \mathbb{R}^d), \int_\Omega u\,dx = \mathcal{L}^N(\Omega)(\theta a + (1-\theta)b)\right\},$$

where $\{W = 0\} = \{a, b\}$, and with corresponding family of singular perturbations (see [10, 12, 13, 22, 25, 28, 29, 30, 32, 33])

$$\int_\Omega (W(u) + \varepsilon^2|\nabla u|^2)\,dx,$$

we attempt to resolve the lack of uniqueness by considering higher gradient penalizations. This is in agreement with higher strain gradient theories in plasticity. To this end, for any open set $A \subset \Omega$ we introduce the family

$$J_\varepsilon(u; A) := \int_A (W(\nabla u) + \varepsilon^2|\nabla^2 u|)\,dx.$$

The characterization of the $\Gamma(L^1)$-limit of these functionals, and, in particular, of the asymptotic behavior of minimizers as $\varepsilon \to 0^+$, is work under progress by Fonseca and Tartar [21]. Here the main difficulties are, essentially, the need to use intrinsically vectorial techniques and the proof of the locality of the $\Gamma(L^1)$-limit. The use of vectorial techniques was successfully exploited in the variational study of the eikonal equation, seen as a partial differential constraint on finite limiting energy fields when in J_ε the density W is allowed to vanish on the sphere (see [5, 8, 9, 18, 26, 27, 24]). In the attempts to ascertain locality, the problems encountered seem to stem from the higher order derivative in the model. Precisely, if we knew that the subadditivity property

$$\Gamma(L^1) - \lim_{\varepsilon\to 0^+} J_\varepsilon(u; A) \le \Gamma(L^1) - \lim_{\varepsilon\to 0^+} J_\varepsilon(u; B) + \Gamma(L^1) - \lim_{\varepsilon\to 0^+} J_\varepsilon(u; A \setminus \overline{C})$$

holds whenever A, B, C are open subsets of Ω with $C \subset\subset B \subset\subset A$, then we would be able to ensure that $\Gamma(L^1) - \lim_{\varepsilon\to 0^+} J_\varepsilon(u;\cdot)$ is a measure, and therefore Radon–Nikodym theorem and a blow-up argument around points on the laminate surfaces would easily yield

$$\Gamma(L^1) - \lim_{\varepsilon\to 0^+} J_\varepsilon(u;\Omega) := \begin{cases} \widehat{\mathbf{m}}\ \mathrm{Per}_\Omega(\{\nabla u = A\}) & \text{if } \nabla u \in BV(\Omega;\{A,B\}), \\ +\infty & \text{otherwise,} \end{cases}$$

for an appropriate surface energy density $\widehat{\mathbf{m}}$. This program may be carried out successfully in the Cahn–Hilliard model where the penalization is of first order. For this reason, and motivated in part by the need to isolate the understanding of the role played by higher order penalizations and the obstacles that they may introduce, we take a step further in the simplification of the original model for J_ε, and we consider the family $\mathcal{F}_\varepsilon$ as defined above.

We note that higher order perturbations of nonconvex problems have been studied recently within the framework of free discontinuity problems. In particular, elliptic regularizations with second order terms were proposed for the approximation of free discontinuity problems related to the Mumford–Shah model for image segmentation in computer vision (see, e.g., [3, 4, 15, 14]).

The one-dimensional problem encapsules the main features of the model. Indeed the relevant contributions of this paper may be found in the next section where, using a priori bounds provided by Gagliardo and Nirenberg inequalities, we are able to show that the limiting energy minimizers are two-phase fields with minimal interfacial perimeter. Further, the resulting interfacial energy per unit area, $\mathbf{m}$, may be computed explicitly as the solution of an auxiliary minimization problem, corresponding to the one-dimensional energetically efficient profiles which connect a at $-\infty$ to b at $+\infty$. The extension of these results to the N-dimensional case follows a standard slicing argument that enables us to reduce it to the one-dimensional setting.

2. The one-dimensional problem.

Let $W : \mathbb{R}^d \to \mathbb{R}$ be a continuous function satisfying the following hypotheses:
(H1) $W(u) = 0$ if and only if $u \in \{a, b\}$;
(H2) there exist constants $C > 0$, $R > \max\{|a|, |b|\}$, such that if $|u| > R$, then $W(u) \geq C|u| - 1/C$.

Let $I := (\alpha, \beta)$ be a fixed open interval in $\mathbb{R}$. Consider the family of functionals indexed by the parameter $\varepsilon > 0$, and defined as

$$\mathcal{F}_\varepsilon(u) := \begin{cases} \int_I \left(\frac{W(u)}{\varepsilon} + \varepsilon^3 |u''|^2\right) dt & \text{if } u \in W^{2,2}(I;\mathbb{R}^d), \\ +\infty & \text{if } u \in L^1(I;\mathbb{R}^d) \setminus W^{2,2}(I;\mathbb{R}^d). \end{cases}$$

We seek to identify the limiting states corresponding to sequences of minimizers for $\mathcal{F}_\varepsilon(\cdot)$, and to this purpose we will use the notion of $\Gamma(L^1)$-convergence. We recall some basic notions on $\Gamma(L^1)$-convergence (for a detailed, comprehensive study we refer the reader to [17]). Let Ω be an open, bounded subset of $\mathbb{R}^N$.

DEFINITION 2.1. *Let $F_n : L^1(\Omega;\mathbb{R}^d) \to [-\infty, +\infty]$ and $u \in L^1(\Omega;\mathbb{R}^d)$. We define*

$$\Gamma(L^1) - \liminf F_n(u) := \inf\left\{ \liminf_{n\to\infty} F_n(u_n) : u_n \to u \quad \text{in } L^1(\Omega;\mathbb{R}^d)\right\}$$

and

$$\Gamma(L^1) - \limsup F_n(u) := \inf\left\{ \limsup_{n\to\infty} F_n(u_n) : u_n \to u \quad \text{in } L^1(\Omega;\mathbb{R}^d)\right\}.$$

If $\Gamma(L^1) - \liminf F_n(u) = \Gamma(L^1) - \limsup F_n(u)$, *then the common value is called the* $\Gamma(L^1)$*-limit of* F_n *at* u, *and is denoted by* $\Gamma(L^1) - \lim F_n(u)$.

Moreover, given a family $F_\varepsilon : L^1(\Omega;\mathbb{R}^d) \to [-\infty,+\infty]$, $\varepsilon > 0$, *if* $u \in L^1(\Omega;\mathbb{R}^d)$, *then we say that* $\Gamma(L^1) - \lim F_\varepsilon(u) = F(u)$ *if* $F(u) = \Gamma(L^1) - \lim F_{\varepsilon_n}(u)$ *for every sequence* $\varepsilon_n \to 0^+$.

It can be shown that $F(u) = \Gamma(L^1)$-limit of F_ε at u if and only if

(i) for every sequences $\{u_n\}$ and $\{\varepsilon_n\}$ such that $u_n \to u$ in $L^1(\Omega;\mathbb{R}^d)$ and $\varepsilon_n \to 0^+$

$$F(u) \le \liminf_{n\to\infty} F_{\varepsilon_n}(u_n);$$

(ii) for every sequence $\{\varepsilon_n\}$ converging to 0^+ there exists a sequence $\{u_n\}$ such that $u_n \to u$ in $L^1(\Omega;\mathbb{R}^d)$ and

$$F(u) = \lim_{n\to\infty} F_{\varepsilon_n}(u_n).$$

In what follows C denotes a generic positive constant which may vary from one formula to the next and from line to line. Also, $\mathcal{L}^N$ stands for the Lebesgue measure in $\mathbb{R}^N$, and $B(x,\delta)$ is the ball centered at the point x and with radius $\delta > 0$.

Define

$$\mathcal{A} := \left\{ f \in W^{2,2}_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^d) : f(t) = b \text{ if } t > C, f(t) = a \text{ if } t < -C \ \text{ for some } \ C > 0 \right\}$$

and

$$\mathbf{m} := \inf\left\{ \int_{\mathbb{R}} (W(f) + |f''|^2)\,dt : f \in \mathcal{A} \right\}. \tag{2.1}$$

The main theorem of this section is the following.

THEOREM 2.2. *For every* $u \in L^1(I;\mathbb{R}^d)$

$$\Gamma(L^1) - \lim_{\varepsilon\to 0^+} \mathcal{F}_\varepsilon(u) = \begin{cases} \mathbf{m}\,\mathrm{Per}_I(\{u = a\}) & \textit{if } u \in BV(I;\{a,b\}), \\ +\infty & \textit{otherwise}. \end{cases}$$

Compactness for energy bounded sequences will rely heavily on the following interpolation inequality due to Gagliardo [23] and Nirenberg [31].

PROPOSITION 2.3. *Let* Ω *be a bounded, open, Lipschitz subset of* $\mathbb{R}^N$. *If* $u \in L^1(\Omega;\mathbb{R}^d)$ *and* $\nabla^2 u \in L^2(\Omega;\mathbb{R}^d)$, *then* $u \in W^{2,2}(\Omega;\mathbb{R}^d)$ *and*

$$\|\nabla u\|_{L^{4/3}} \le C\left(\|u\|_{L^1}^{1/2} \|\nabla^2 u\|_{L^2}^{1/2} + \|u\|_{L^1} \right), \tag{2.2}$$

where $C = C(\Omega, N, d)$.

In what follows we will also use the following interpolation inequality.

LEMMA 2.4. *Let* $\varphi : (0,+\infty) \to \mathbb{R}$ *be a convex, nondecreasing function in* $\mathbb{R}^+$, *and let* J *be* $\mathbb{R}$ *or a half-line. Then for every function* $u \in L^1_{\mathrm{loc}}(J;\mathbb{R}^d)$ *with* $u'' \in L^1_{\mathrm{loc}}(J;\mathbb{R}^d)$ *we have*

$$\int_J \varphi\left(\frac{|u'|}{4d}\right)\,dt \le \frac{3}{4}\int_J [\varphi(|u|) + \varphi(|u''|)]\,dt. \tag{2.3}$$

Proof. The case where $\varphi(t) = |t|^p$ may be found in Adams [1, Lemma 4.10]. First consider the real valued case where $d = 1$. Given $u \in W^{2,1}(0,1)$, fix $\theta \in (0, 1/3)$ and $\eta \in (2/3, 1)$, and by virtue of the mean value theorem, find $\xi \in (0,1)$ such that

$$u'(\xi) = \frac{u(\theta) - u(\eta)}{\theta - \eta};$$

hence,

$$|u'(x)| \leq \frac{|u(\theta) - u(\eta)|}{|\theta - \eta|} + \int_\xi^x |u''|\, dt \leq \frac{|u(\theta) - u(\eta)|}{|\theta - \eta|} + \int_0^1 |u''|\, dt$$

for all $x \in (0,1)$, which, by the choice of θ and η, implies that

$$|u'(x)| \leq 3|u(\theta)| + 3|u(\eta)| + \int_0^1 |u''|\, dt \quad \text{for all } x \in (0,1).$$

Integrating in θ and η and multiplying both sides by 9 we get

$$\begin{aligned} |u'(x)| &\leq 3\int_0^{1/3} |u|\, dt + 3\int_{2/3}^1 |u|\, dt + \int_0^1 |u''|\, dt \\ &\leq 3\int_0^1 |u|\, dt + \int_0^1 |u''|\, dt \end{aligned}$$

for all $x \in (0,1)$. Now, dividing both sides by 4, and using the convexity and monotonicity properties of φ, together with Jensen's inequality, we obtain

$$\begin{aligned} \varphi\left(\frac{|u'(x)|}{4}\right) &\leq \varphi\left(\frac{3}{4}\int_0^1 |u|\, dt + \frac{1}{4}\int_0^1 |u''|\, dt\right) \\ &\leq \frac{3}{4}\varphi\left(\int_0^1 |u|\, dt\right) + \frac{1}{4}\varphi\left(\int_0^1 |u''|\, dt\right) \\ &\leq \frac{3}{4}\int_0^1 [\varphi(|u|) + \varphi(|u''|)]\, dt \end{aligned}$$

for all $x \in (0,1)$.

Finally, integrating in x we have

$$\int_0^1 \varphi\left(\frac{|u'|}{4}\right) dt \leq \frac{3}{4}\int_0^1 [\varphi(|u|) + \varphi(|u''|)]\, dt\,.$$

Dividing J in disjoint intervals of length 1 and applying this argument to each one of them we conclude that

$$\int_J \varphi\left(\frac{|u'|}{4}\right) dt \leq \frac{3}{4}\int_J [\varphi(|u|) + \varphi(|u''|)]\, dt\,.$$

If u takes values in $\mathbb{R}^d$, $d \geq 2$, then

$$\begin{aligned}\int_J \varphi\left(\frac{|u'|}{4d}\right) dt &\leq \int_J \varphi\left(\sum_{i=1}^d \frac{|u_i'|}{4d}\right) dt\\ &\leq \frac{1}{d}\sum_{i=1}^d \int_J \varphi\left(\frac{|u_i'|}{4}\right) dt\\ &\leq \frac{3}{4d}\sum_{i=1}^d \int_J [\varphi(|u_i|)+\varphi(|u_i''|)]\, dt\\ &\leq \frac{3}{4d}\sum_{i=1}^d \int_J [\varphi(|u|)+\varphi(|u''|)]\, dt\\ &= \frac{3}{4}\int_J [\varphi(|u|)+\varphi(|u''|)]\, dt,\end{aligned}$$

which proves the lemma. □

In that which follows we will exploit the auxiliary functions $G, H : \mathbb{R}^{2d} \to \mathbb{R}$, which take into account the energy stored on an interfacial layer

$$G(w,z) := \inf\left\{ \int_0^1 (W(g) + |g''|^2)\, dt : \; g \in C^2([0,1];\mathbb{R}^d),\, g(0) = w,\, g(1) = b,\right.$$
$$\left. g'(0) = z,\, g'(1) = 0 \right\},$$

$$H(w,z) = \inf\left\{ \int_0^1 (W(h) + |h''|^2)\, dt : \; h \in C^2([0,1];\mathbb{R}^d),\, h(0) = a,\, h(1) = w,\right.$$
$$\left. h'(0) = 0,\, h'(1) = z \right\}.$$

Testing G and H with third degree polynomials g and h, respectively, satisfying the boundary conditions, it can be shown that

$$\lim_{(w,z)\to(b,0)} G(w,z) = 0\,, \qquad \lim_{(w,z)\to(a,0)} H(w,z) = 0\,. \tag{2.4}$$

LEMMA 2.5. *The constant* $\mathbf{m}$ *is positive and*

$$\mathbf{m} = \min\left\{ \int_{\mathbb{R}} (W(f) + |f''|^2)\, dt : f \in W^{2,2}_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^d),\, \lim_{t\to+\infty} f(t) = b,\, \lim_{t\to-\infty} f(t) = a \right\}.$$

Proof. Step 1. We start by proving that $\mathbf{m} > 0$. Suppose that $\mathbf{m} = 0$ and let $\{f_n\}$ be a minimizing sequence of admissible functions in $\mathcal{A}$. Let

$$S := \left\{ x \in \mathbb{R}^d : |x - a| = \frac{|b-a|}{2} \right\}.$$

By the Sobolev embedding theorem each function f_n belongs to $C^1(\mathbb{R};\mathbb{R}^d)$, and since $f_n(t) = b$ for $t > M_n$ and $f_n(t) = a$ if $t < -M_n$ for a suitable $M_n > 0$, there must exist a point $t_n \in \mathbb{R}$ such that $f_n(t_n) \in S$. By performing a simple translation in the variable, with no loss of generality, we may assume that $f_n(0) \in S$. As $\mathbf{m} = 0$, we have that $\|f_n''\|_{L^2} \to 0$; moreover, by (H2), fixed a bounded interval $J \subset \mathbb{R}$ containing the origin, $\{f_n\}$ is equibounded in $L^1(J;\mathbb{R}^d)$, and Proposition 2.3 implies that $\{f_n'\}$ is equibounded in $L^{4/3}(J;\mathbb{R}^d)$. Therefore, by the Sobolev embedding theorem it follows that $\{f_n\}$ is bounded in $W^{2,2}(J;\mathbb{R}^d)$, and we may extract a subsequence $f_{n_i}|_J$ of restricted functions converging in $W^{1,\infty}$ to an affine function $f : J \to \mathbb{R}^d$ such that $f(0) =: c \in S$. Setting $f(t) := c + tv$ for some $v \in \mathbb{R}^d$, we have

$$\begin{aligned}\mathbf{m} &= \lim_{n\to\infty} \int_{\mathbb{R}} (W(f_{n_i}) + |f_{n_i}''|^2)\, dt \\ &\geq \lim_{n\to\infty} \int_J (W(f_{n_i}) + |f_{n_i}''|^2)\, dt \\ &\geq \int_J W(c+tv)\, dt > 0\,,\end{aligned}$$

because if $\int_J W(c+tv)\, dt = 0$, then $c + tv$ should belong to $\{a, b\}$ for all $t \in J$, and therefore $v = 0$ and $c \in \{a, b\}$, which is not possible since $c \in S$. We arrived at a contradiction, and thus $\mathbf{m} > 0$.

Step 2. Next we prove that $\mathbf{m} = \widetilde{\mathbf{m}}$, where

(2.5)
$$\widetilde{\mathbf{m}} := \inf\left\{ \int_{\mathbb{R}} (W(f) + |f''|^2)\, dt : f \in W^{2,2}_{\text{loc}}(\mathbb{R};\mathbb{R}^d),\ \lim_{t\to+\infty} f(t) = b,\ \lim_{t\to-\infty} f(t) = a \right\}.$$

It is clear that $\mathbf{m} \geq \widetilde{\mathbf{m}}$.

Conversely, fix $\delta > 0$ and let f be a function admissible for $\widetilde{\mathbf{m}}$ and such that

$$\widetilde{\mathbf{m}} + \delta \geq \int_{\mathbb{R}} (W(f) + |f''|^2)\, dt.$$

We claim that we may find two sequences $\{x_i\}$ and $\{y_i\}$ converging to $+\infty$ and $-\infty$, respectively, such that

$$|f'(x_i)| + |f'(y_i)| + |f(x_i) - b| + |f(y_i) - a| \to 0 \tag{2.6}$$

as $i \to \infty$. Indeed, fix $\tau < |b-a|/2$ and consider a convex, nondecreasing function $\varphi : \mathbb{R} \to [0, +\infty)$ such that $\varphi(t) \leq t^2$ for every $t \in \mathbb{R}$, $\varphi(|y|) \leq W(y+b)$ for every $y \in B(0,\tau) \subset \mathbb{R}^d$, and $\varphi(t) = 0$ if and only if $t = 0$. To prove the existence of φ it suffices to set

$$\begin{aligned}\varphi(t) := \sup\{g : \mathbb{R} \to [0,+\infty) : g \text{ is convex, nondecreasing,}\\ g(t) \leq t^2 \text{ for all } t \in \mathbb{R},\ g(|y|) \leq W(y+b) \text{ for all } y \in B(0,\tau)\}\end{aligned}$$

and use hypothesis (H1). Let $R > 0$ be such that $|f(t) - b| < \tau$ whenever $t > R$. Applying Lemma 2.4 to the function $f - b$, and using the properties of the function φ, we obtain

$$\begin{aligned}\int_R^{+\infty} \varphi\left(\frac{|f'|}{4d}\right) dt &\leq \frac{3}{4} \int_R^{+\infty} [\varphi(|f-b|) + \varphi(|f''|)]\, dt \\ &\leq \frac{3}{4} \int_R^{+\infty} (W(f) + |f''|^2)\, dt \leq \frac{3\,(\widetilde{\mathbf{m}} + \delta)}{4}\,.\end{aligned}$$

Thus $\varphi\left(\frac{|f'|}{4d}\right)$ is integrable on $[R,+\infty)$, and so there exists a sequence of points $x_n \to +\infty$ such that $\lim_{n\to\infty}\varphi\,\left(\frac{|f'(x_n)|}{4d}\right)$=0, and since φ is monotone nondecreasing on $[0,+\infty)$, with $\varphi(t)=0$ if and only if $t=0$, we conclude that $\lim_{n\to\infty} f'(x_n)=0$. Repeating this argument with the point a in place of b, we are now in a position to assert the existence of two sequences satisfying (2.6).

Set

$$g_i(t) := g(t-x_i), \qquad h_i(t) := h(t-y_i+1),$$

where g and h are admissible for G and H, respectively, and

$$\int_0^1 (W(g)+|g''|^2)\,dt \le G(f(x_i),f'(x_i))+\delta,$$

$$\int_0^1 (W(h)+|h''|^2)\,dt \le H(f(y_i),f'(y_i))+\delta.$$

We define

$$\widetilde{f}_i(t) := \begin{cases} b & \text{if } t \ge x_i+1,\\ g_i(t) & \text{if } t\in[x_i,x_i+1],\\ f(t) & \text{if } t\in[y_i,x_i],\\ h_i(t) & \text{if } t\in[y_i-1,y_i],\\ a & \text{if } t\le y_i-1.\end{cases}$$

Clearly $\widetilde{f}_i$ is admissible for $\mathbf{m}$, and we have

$$\begin{aligned}\widetilde{\mathbf{m}}+\delta &\ge \int_{\mathbb{R}}(W(f)+|f''|^2)\,dt \ge \int_{y_i}^{x_i}(W(f)+|f''|^2)\,dt\\ &= \int_{\mathbb{R}}(W(\widetilde{f}_i)+|\widetilde{f}_i''|^2)\,dt - \int_{x_i}^{x_i+1}(W(g_i)+|g_i''|^2)\,dt\\ &\quad -\int_{y_i-1}^{y_i}(W(h_i)+|h_i''|^2)\,dt\\ &\ge \mathbf{m}-G(f(x_i),f'(x_i))-H(f(y_i),f'(y_i))-2\delta.\end{aligned}$$

The inequality $\widetilde{\mathbf{m}}\ge\mathbf{m}$ now follows by letting $\delta\to 0^+$, $i\to\infty$, and using (2.4).

Step 3. Finally, we prove that $\mathbf{m}$ is attained, or, equivalently, that $\widetilde{\mathbf{m}}$ admits a minimizer. Let $\{f_n\}$ be a minimizing sequence for $\widetilde{\mathbf{m}}$. Possibly passing to a subsequence, and making a translation change of variables, we may assume as before that $f_n(0)\in S$, where S was defined in Step 1, and that the sequence of C^1 functions $\{f_n\}$ converges in $W^{1,\infty}_{\text{loc}}$ to a C^1 function $f:\mathbb{R}\to\mathbb{R}^d$. If the function f is admissible, then it realizes the infimum, since

$$\int_{\mathbb{R}}(W(f)+|f''|^2)\,dt \le \varliminf_{n\to\infty}\int_{\mathbb{R}}(W(f_n)+|f_n''|^2)\,dt,$$

where we have used Fatou's lemma and the lower semicontinuity of the L^2 norm of the second derivative. In order to prove that f approaches a and b at infinity, set

$$L := \left\{ l\in\mathbb{R}^d \mid \ l \text{ is a limit point of } f(t) \text{ when } t\to+\infty\right\}.$$

The integrability of $W(f)$ and (H1) imply that a or b must belong to L. Suppose that $b \in L$, and that there is another limiting value $l \in L$. Note that, without loss of generality, we may assume that $l \neq a$, for if $l = a$, then, by the continuity of f, there would exist a sequence $y_i \to +\infty$ such that $f(y_i) \in S$; hence, for a subsequence (not relabelled) $f(y_i) \to l' \in S$. Consider two monotone sequences of points $\{x_i\}$ and $\{z_i\}$ such that $x_{i+1} - x_i \geq 3$, $z_i \in [x_i + 1, x_{i+1} - 1]$, $f(x_i) \to b$ and $f(z_i) \to l$, and for $0 < \delta < \min\{|l-a|, |l-b|\}$ we introduce still another constant $\widehat{\mathbf{m}}$ defined as follows:

$$\widehat{\mathbf{m}} := \inf\left\{ \int_x^y (W(g) + |g''|^2)\, dt : y - x \geq 3,\ z \in [x+1, y-1], \right. \\ \left. g \in W^{2,2}((x,y);\mathbb{R}^d),\ |g(z) - l| \leq \delta \right\}.$$

We claim that $\widehat{\mathbf{m}} = 0$. Indeed, if $\widehat{\mathbf{m}} > 0$, then there would exist n_0 such that, for $n \geq n_0$, $|f(z_n) - l| \leq \delta$, and it would follow that

$$\begin{aligned}\int_{\mathbb{R}} (W(f) + |f''|^2)\, dt &= \int_{-\infty}^{x_{n_0}} (W(f) + |f''|^2)\, dt + \sum_{i=n_0}^{\infty} \int_{x_i}^{x_{i+1}} (W(f) + |f''|^2)\, dt \\ &\geq \int_{-\infty}^{x_{n_0}} (W(f) + |f''|^2)\, dt + \sum_{i=n_0}^{\infty} \widehat{\mathbf{m}} = +\infty.\end{aligned}$$

On the other hand, we can show that the assertion $\widehat{\mathbf{m}} = 0$ leads to a contradiction. The reasoning is similar to the one used in Step 1 for the constant $\mathbf{m}$. Let $g_n \in W^{2,2}((x_n, y_n);\mathbb{R}^d)$ minimize $\widehat{\mathbf{m}}$. Translating the intervals, without loss of generality we can suppose that $z_n = 0$, thus $x_n \leq -1$ and $y_n \geq 1$, and possibly passing to a subsequence (not relabelled), we may assume that the functions g_n converge in $W^{1,\infty}([-1,1];\mathbb{R}^d)$ to an affine function $g(t) = d + tv$. Therefore

$$\begin{aligned}\widehat{\mathbf{m}} &\geq \lim_{n\to\infty} \int_{x_n}^{y_n} (W(g_n) + |g_n''|^2)\, dt \\ &\geq \lim_{n\to\infty} \int_{-1}^{1} (W(g_n) + |g_n''|^2)\, dt \\ &\geq \int_{-1}^{1} W(d + tv)\, dt > 0\,,\end{aligned}$$

because if $\int_{-1}^{1} W(d+tv)\, dt = 0$, then $d + tv$ should belong to $\{a, b\}$ for all $t \in (-1, 1)$, i.e., $v = 0$ and $d \in \{a, b\}$. This is not possible since $g_n(0) \to d$, $g_n(0) \in B(l,\delta)$, and $a, b \notin B(l, \delta)$. We conclude that $f(t) \to b$ as $t \to +\infty$.

Similarly, if $a \in L$, then $f(t)$ converges to a as $t \to -\infty$.

If the limits of f at $+\infty$ and $-\infty$ are, respectively, a and b, then $f(-t)$ is still a minimizer and it converges to b and a at, respectively, $+\infty$ and $-\infty$.

It remains to exclude the possibility that the two limits coincide. Suppose that $\lim_{t\to\pm\infty} f(t) = a$. As in Step 2, by virtue of Lemma 2.4 we can find a sequence of points $x_n \to +\infty$ such that

$$|f'(x_n)| + |f(x_n) - a| \to 0$$

and, due to the convergence of f_n to f in $W^{1,\infty}_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^d)$, there exists a subsequence $\{f_{k_n}\}$ such that $f_{k_n}(x_n) \to a$ and $f'_{k_n}(x_n) \to 0$. Hence, we have

$$\begin{aligned}\int_{\mathbb{R}} (W(f_{k_n}) + |f''_{k_n}|^2)\,dt &= \int_{-\infty}^{x_n} (W(f_{k_n}) + |f''_{k_n}|^2)\,dt + \int_{x_n}^{+\infty} (W(f_{k_n}) + |f''_{k_n}|^2)\,dt \\ &\geq \int_{-\infty}^{x_n} (W(f_{k_n}) + |f''_{k_n}|^2)\,dt + \widetilde{\mathbf{m}} - H(f_{k_n}(x_n), f'_{k_n}(x_n))\,,\end{aligned}$$

and letting $n \to \infty$ we deduce that

$$\begin{aligned}\widetilde{\mathbf{m}} &\geq \limsup_{n\to\infty} \int_{-\infty}^{x_n} (W(f_{k_n}) + |f''_{k_n}|^2)\,dt + \widetilde{\mathbf{m}} \\ &\geq \limsup_{n\to\infty} \int_{-\infty}^{x_n} W(f_{k_n})\,dt + \widetilde{\mathbf{m}} \\ &= \int_{\mathbb{R}} W(f)\,dt + \widetilde{\mathbf{m}}\,.\end{aligned}$$

This would imply that f is constantly equal to a, but since $f_n(0) \in S$ for every n we also have $f(0) = \lim_{n\to\infty} f_n(0) \in S$ which is in contradiction with $a \notin S$. The case where $\lim_{t\to\pm\infty} f(t) = b$ is treated in an analogous way. □

Remark 2.6. A simple rescaling argument provides equipartition of energy. Precisely, if f realizes the minimum $\mathbf{m}$, then

$$\int_{\mathbb{R}} W(f)\,dt = 3\int_{\mathbb{R}} |f''|^2\,dt\,.$$

It suffices to use the fact that

$$\int_{\mathbb{R}} (W(f) + |f''|^2)\,dt \leq \int_{\mathbb{R}} (W(f_\lambda) + |f''_\lambda|^2)\,dt$$

for all $\lambda > 0$, where $f_\lambda(x) := f(\lambda x)$.

We now state and prove the compactness result for sequences with finite energy.

Proposition 2.7. *If $u_\varepsilon \in W^{2,2}(I;\mathbb{R}^d)$ satisfy $\liminf_{\varepsilon\to 0^+} \mathcal{F}_\varepsilon(u_\varepsilon) < +\infty$, then there exists a subsequence $\{u_{\varepsilon_n}\} \subset \{u_\varepsilon\}$ and $u \in BV(I;\{a,b\})$ such that $u_{\varepsilon_n} \to u$ in $L^1(I;\mathbb{R}^d)$. Moreover,*

$$\liminf_{\varepsilon\to 0^+} \mathcal{F}_\varepsilon(u_\varepsilon) \geq \mathbf{m}\,\mathrm{Per}_I(\{u = a\}).$$

Proof. Suppose that $\liminf_{\varepsilon\to 0^+} \mathcal{F}_\varepsilon(u_\varepsilon) =: K < +\infty$. We claim that there exists a function $u \in BV(I;\{a,b\})$ such that, up to a subsequence, $u_\varepsilon \to u$. Extract a subsequence from the start (not relabelled) realizing $\liminf_{\varepsilon\to 0^+} \mathcal{F}_\varepsilon(u_\varepsilon)$. We have

$$W(u_\varepsilon) \to 0 \text{ in } L^1, \quad ||u''||_{L^2} \leq C\varepsilon^{-3/2}.$$

By (H2)

$$||u_\varepsilon||_{L^1} \leq R\,\mathcal{L}^1(I) + \int_{\{|u_\varepsilon|>R\}} \left(\frac{1}{C^2} + \frac{1}{C}W(u_\varepsilon)\right)\,dt \leq \widetilde{C};$$

therefore, by the Gagliardo and Nirenberg inequality (2.2) we conclude that

$$||u'_\varepsilon||_{L^{4/3}} \leq C(||u_\varepsilon||^{1/2}_{L^1}||u''_\varepsilon||^{1/2}_{L^2} + ||u_\varepsilon||_{L^1}) \leq \widetilde{C}\varepsilon^{-3/4}. \tag{2.7}$$

Also

$$\mathcal{L}^1(\{|u_\varepsilon| > R\}) \to 0 \text{ as } \varepsilon \to 0^+. \tag{2.8}$$

Indeed, if $\tau := \inf\{W(\xi) : |\xi| > R\}$, then by (H1) we have $\tau > 0$, and therefore (2.8) follows from the fact that

$$0 = \lim_{\varepsilon\to 0^+} \int_I W(u_\varepsilon)\, dt \geq \limsup_{\varepsilon\to 0^+} \int_{\{|u_\varepsilon|>R\}} W(u_\varepsilon)\, dt \geq \tau \limsup_{\varepsilon\to 0^+} \mathcal{L}^1(\{|u_\varepsilon| > R\}).$$

Since $\{u_\varepsilon\}$ is bounded in L^1, we may extract a further subsequence (not relabelled) generating a Young measure $\{\nu_t\}_{t\in I}$. In particular, if $f : I \times \mathbb{R}^d \to [0, +\infty)$ is a Carathéodory function such that $\{f(\cdot, u_\varepsilon(\cdot))\}$ is equi-integrable, then $f(\cdot, u_\varepsilon(\cdot)) \rightharpoonup \bar{f}$ in L^1 where (see [11, 34])

$$\bar{f}(t) := \int_{\mathbb{R}} f(t, y) d\nu_t(y), \quad \text{almost everywhere (a.e.) } t \in I.$$

Setting $f(y) := \min\{W(y), 1\}$, it follows that

$$0 = \lim \int_I f(u_\varepsilon)\, dt = \int_I \int_{\mathbb{R}} f(y) d\nu_t(y)\, dt;$$

hence, since $f(y) = 0$ if and only $y \in \{a, b\}$, we have

$$\nu_t = \theta(t)\delta_{y=a} + (1 - \theta(t))\delta_{y=b}$$

for some $\theta \in L^\infty(I, [0, 1])$. We claim that

$$\theta \in \{0, 1\} \text{ a.e. in } I, \quad \text{i.e., } \theta = \chi_E \text{ for some measurable subset } E \subset I. \tag{2.9}$$

Suppose that the claim holds. Define

$$u(t) := a\chi_E(t) + b(1 - \chi_E(t)).$$

Then $u_\varepsilon \to u$ strongly in L^1. Indeed, let

$$\varphi(y) := \begin{cases} R\frac{y}{|y|} & \text{if } |y| > R, \\ y & \text{if } |y| \leq R. \end{cases}$$

Note that $u = \varphi(u)$, and recall that $W(y) \geq C|y| - 1/C$ if $|y| > R$, with $R > \max\{|a|, |b|\}$. Then

$$\begin{aligned} \int_I |u_\varepsilon - u|\, dt &\leq \int_I |\varphi(u_\varepsilon) - u|\, dt + 2\int_{|u_\varepsilon|>R} |u_\varepsilon|\, dt \\ &\leq \int_I |\varphi(u_\varepsilon) - u|\, dt + \frac{2}{C}\int_I W(u_\varepsilon)\, dt + \frac{2}{C^2}\mathcal{L}^1(\{|u_\varepsilon| > R\}). \end{aligned}$$

Therefore, by (2.8) and (2.9), we conclude that

$$\begin{aligned} \lim_{\varepsilon\to 0^+} \int_I |u_\varepsilon - u|\, dt &= \int_I \int_{\mathbb{R}} |\varphi(y) - u(t)| d\nu_t(y) dt \\ &= \int_E |\varphi(a) - u(t)|\, dt + \int_{I\setminus E} |\varphi(b) - u(t)|\, dt \\ &= \int_E |a - u(t)|\, dt + \int_{I\setminus E} |b - u(t)|\, dt = 0. \end{aligned}$$

To prove (2.9), define

$$X := \left\{ t \in I : \frac{1}{|B(t,\delta)|} \int_{B(t,\delta)} \theta(s)ds \in (0,1) \ \text{ for all } \ \delta > 0 \right\}.$$

We show that the cardinality of X (call it L) cannot exceed the integer part of $K/\mathbf{m}$; hence $\theta \in \{0,1\}$ a.e. and $u \in BV(I;\{a,b\})$ with $\mathbf{m}\,\mathrm{Per}_I(\{u=a\}) = \mathbf{m}\,L \le K$ which gives the result.

Indeed, suppose that there were l distinct points of I in X, $s_1 < s_2 < \cdots < s_l$. Let $\delta_0 := \min\{|s_i - s_{i+1}| : i = 1, \dots, l-1\}$. Choose $\delta_1 < \delta_0/2$ such that for all $\delta \le \delta_1$, and all $i \in \{1, \dots, l\}$,

$$\int_{B(s_i,\delta)} \theta(s)ds > 0, \qquad \int_{B(s_i,\delta)} (1-\theta(s))ds > 0.$$

Fix $0 < \eta < |b-a|/2$, let φ_η be a cut-off function with support on $B(a,\eta)$, $\varphi_\eta(a) = 1$, ψ_η is a cut-off function with support on $B(0,\eta)$, $\psi_\eta(0) = 1$, and γ_η is a cut-off function with support on $B(b,\eta)$, $\gamma_\eta(b) = 1$. By (2.7) $\psi_\eta(\varepsilon u'_\varepsilon)$ converges strongly to $\psi_\eta(0)$ in L^1, $\varphi_\eta(u_\varepsilon)$ converges in L^∞ weak-* to $\theta\varphi_\eta(a) + (1-\theta)\varphi_\eta(b)$, and we have

$$\begin{aligned}\lim_{\varepsilon\to 0^+} \int_{B(s_i,\delta_1)} \psi_\eta(\varepsilon u'_\varepsilon)\varphi_\eta(u_\varepsilon)dt &= \int_{B(s_i,\delta_1)} \psi_\eta(0)[\theta(t)\varphi_\eta(a) + (1-\theta(t))\varphi_\eta(b)]dt \\ &= \int_{B(s_i,\delta_1)} \theta(t)\,dt > 0,\end{aligned}$$

and, similarly,

$$\lim_{\varepsilon\to 0^+} \int_{B(s_i,\delta_1)} \psi_\eta(\varepsilon u'_\varepsilon)\gamma_\eta(u_\varepsilon)\,dt = \int_{B(s_i,\delta_1)} (1-\theta(t))\,dt > 0\,.$$

Thus, for each $i \in \{1, \dots, l\}$ and each $\varepsilon > 0$, we may find $x^+_{\varepsilon,i}, x^-_{\varepsilon,i} \in B(s_i,\delta_1)$ such that

$$u_\varepsilon(x^+_{\varepsilon,i}) \in B(b,\eta), u_\varepsilon(x^-_{\varepsilon,i}) \in B(a,\eta), |\varepsilon u'_\varepsilon(x^+_{\varepsilon,i})| < \eta, |\varepsilon u'_\varepsilon(x^-_{\varepsilon,i})| < \eta\,. \tag{2.10}$$

Define

$$g_{\varepsilon,i}(t) := \widehat{g_{\varepsilon,i}}\left(t - \frac{x^+_{\varepsilon,i}}{\varepsilon}\right), \quad h_{\varepsilon,i}(t) := \widehat{h_{\varepsilon,i}}\left(t - \frac{x^-_{\varepsilon,i}}{\varepsilon} + 1\right),$$

where the functions $\widehat{g_{\varepsilon,i}}$ and $\widehat{h_{\varepsilon,i}}$ are admissible for $G(u_\varepsilon(x^+_{\varepsilon,i}), \varepsilon u'_\varepsilon(x^+_{\varepsilon,i}))$ and for $H(u_\varepsilon(x^-_{\varepsilon,i}), \varepsilon u'_\varepsilon(x^-_{\varepsilon,i}))$, respectively, with

$$\int_0^1 (W(\widehat{g_{\varepsilon,i}}) + |\widehat{g_{\varepsilon,i}}''|^2)\,dt \le G(u_\varepsilon(x^+_{\varepsilon,i}), \varepsilon u'_\varepsilon(x^+_{\varepsilon,i})) + \varepsilon,$$

and

$$\int_0^1 \left(W(\widehat{h_{\varepsilon,i}}) + |\widehat{h_{\varepsilon,i}}''|^2\right)\,dt \le H(u_\varepsilon(x^-_{\varepsilon,i}), \varepsilon u'_\varepsilon(x^-_{\varepsilon,i}))) + \varepsilon.$$

Construct the functions

$$v_{\varepsilon,i}(t) := \begin{cases} b & \text{if } \ t \geq \frac{x^+_{\varepsilon,i}}{\varepsilon} + 1, \\[2ex] g_{\varepsilon,i}(t) & \text{if } t \in \left[\frac{x^+_{\varepsilon,i}}{\varepsilon}, \frac{x^+_{\varepsilon,i}}{\varepsilon} + 1\right], \\[2ex] u_\varepsilon(\varepsilon t) & \text{if } t \in \left[\frac{x^-_{\varepsilon,i}}{\varepsilon}, \frac{x^+_{\varepsilon,i}}{\varepsilon}\right], \\[2ex] h_{\varepsilon,i}(t) & \text{if } t \in \left[\frac{x^+_{\varepsilon,i}}{\varepsilon} - 1, \frac{x^+_{\varepsilon,i}}{\varepsilon}\right], \\[2ex] a & \text{if } \ t \leq \frac{x^+_{\varepsilon,i}}{\varepsilon} - 1. \end{cases}$$

Then $v_{\varepsilon,i}$ are admissible for $\mathbf{m}$, and since the intervals $[x^-_{\varepsilon,i}, x^+_{\varepsilon,i}]$ are disjoint we have

$$\begin{aligned} K &\geq \liminf_{\varepsilon \to 0^+} \sum_{i=1}^{l} \int_{x^-_{\varepsilon,i}}^{x^+_{\varepsilon,i}} \left(\frac{1}{\varepsilon} W(u_\varepsilon) + \varepsilon^3 |u''_\varepsilon|^2\right) dt \\ &= \liminf_{\varepsilon \to 0^+} \sum_{i=1}^{l} \int_{\frac{x^-_{\varepsilon,i}}{\varepsilon}}^{\frac{x^+_{\varepsilon,i}}{\varepsilon}} (W(v_{\varepsilon,i}) + |v''_{\varepsilon,i}|^2)\, dt \\ &\geq \mathbf{m} l - \limsup_{\varepsilon \to 0^+} \sum_{i=1}^{l} [H(u_\varepsilon(x^-_{\varepsilon,i}), \varepsilon u'_\varepsilon(x^-_{\varepsilon,i})) + G(u_\varepsilon(x^+_{\varepsilon,i}), \varepsilon u'_\varepsilon(x^+_{\varepsilon,i}))]. \end{aligned}$$

Letting $\eta \to 0^+$, we conclude that $K \geq \mathbf{m}\, l$, where we have used (2.4) and (2.10). The result now follows from the arbitrariness of $l \leq L$ and the fact that $\mathbf{m} > 0$ (see Lemma 2.5). $\square$

As an immediate consequence of the previous result we have the following.

Corollary 2.8. *If $u \in L^1(I; \mathbb{R}^d)$ and $\Gamma(L^1) - \liminf_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) < +\infty$, then the function u belongs to $BV(I; \{a, b\})$ and*

$$\Gamma(L^1) - \liminf_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) \geq \mathbf{m} \operatorname{Per}_I(\{u = a\}).$$

Now we turn our attention to the $\Gamma(L^1) - \limsup_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon$.

Proposition 2.9. *If $u \in BV(I; \{a, b\})$, then*

$$\Gamma(L^1) - \limsup_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) \leq \mathbf{m} \operatorname{Per}_I(\{u = a\}).$$

Proof. Suppose that $S(u) = \{s_1, \ldots, s_l\} \subset I = (\alpha, \beta)$ is the jump set of the function u, with $\alpha < s_1 < \cdots < s_l < \beta$. Set $\delta_0 := \min\{s_{j+1} - s_j : j = 0, \ldots, l\}$, with $s_0 := \alpha$ and $s_{l+1} := \beta$, and let $I_i := [\frac{s_{i-1}+s_i}{2}, \frac{s_i+s_{i+1}}{2}]$ for $i = 1, \ldots l$. Fix $\delta \in (0, \delta_0)$ and let $f \in \mathcal{A}$ be an admissible function for $\mathbf{m}$, with $f \in W^{2,2}_{\mathrm{loc}}(\mathbb{R}; \mathbb{R}^d)$, $f(t) = b$ if $t > M$, $f(t) = a$ if $t < -M$,

$$\int_{\mathbb{R}} (W(f) + |f''|^2)\, dt \leq \mathbf{m} + \frac{\delta}{l}. \tag{2.11}$$

Consider a sequence $\varepsilon_n \to 0^+$, and choose n sufficiently large so that $\frac{\delta}{2\varepsilon_n} > M$.

Define

$$u_n(t) := \begin{cases} f\left(\frac{t-s_i}{\varepsilon_n}\right) & \text{if } t \in \left[\frac{s_{i-1}+s_i}{2}, \frac{s_{i+1}+s_i}{2}\right] \text{ and if } [u](s_i) = b-a, \\ f\left(-\frac{t-s_i}{\varepsilon_n}\right) & \text{if } t \in \left[\frac{s_{i-1}+s_i}{2}, \frac{s_{i+1}+s_i}{2}\right] \text{ and if } [u](s_i) = a-b, \\ u(t) & \text{otherwise,} \end{cases}$$

where $[u](s_i) := u(s_i) - u(s_{i-1})$ for $i = 1, \dots, l$. We note that $u_n \in W^{2,2}(I;\mathbb{R}^d)$. Indeed, if $[u](s_i) = b-a$, then $u\left(\frac{s_{i-1}+s_i}{2}\right) = a$, $u\left(\frac{s_{i+1}+s_i}{2}\right) = b$, and since

$$\frac{s_{i-1}-s_i}{2\varepsilon_n} < -\frac{\delta}{2\varepsilon_n} < -M, \qquad \frac{s_{i+1}-s_i}{2\varepsilon_n} > \frac{\delta}{2\varepsilon_n} > M,$$

we have that

$$f\left(\frac{s_{i-1}-s_i}{2\varepsilon_n}\right) = a, \quad f\left(\frac{s_{i+1}-s_i}{2\varepsilon_n}\right) = b, \quad f'\left(\frac{s_{i\pm 1}-s_i}{2\varepsilon_n}\right) = 0.$$

A similar argument applies to the case where $[u](s_i) = a-b$.

Since $u_n \to u$ in $L^1(I;\mathbb{R}^d)$ we conclude that

$$\begin{aligned}\lim_{n\to\infty} \mathcal{F}_{\varepsilon_n}(u_n) &= \lim_{n\to\infty} \sum_{i=1}^{l} \int_{I_i} \left(\frac{W(u_n)}{\varepsilon_n} + \varepsilon_n^3 |u_n''|^2\right) dt \\ &= \lim_{n\to\infty} \left\{ \sum_{i=1,\dots,l,\ [u](s_i)=b-a} \int_{\frac{s_{i-1}+s_i}{2\varepsilon_n}}^{\frac{s_{i+1}+s_i}{2\varepsilon_n}} (W(f(t)) + |f''(t)|^2)\, dt \right. \\ &\qquad \left. + \sum_{i=1,\dots,l,\ [u](s_i)=a-b} \int_{\frac{s_{i-1}+s_i}{2\varepsilon_n}}^{\frac{s_{i+1}+s_i}{2\varepsilon_n}} (W(f(-t)) + |f''(-t)|^2)\, dx \right\} \\ &= \left\{ \sum_{i=1,\dots,l,\ [u](s_i)=b-a} \int_{\mathbb{R}} (W(f(t)) + |f''(t)|^2)\, dt \right. \\ &\qquad \left. + \sum_{i=1,\dots,l,\ [u](s_i)=a-b} \int_{\mathbb{R}} (W(f(-t)) + |f''(-t)|^2)\, dt \right\} \\ &= l \int_{\mathbb{R}} (W(f) + |f''|^2)\, dt \\ &\leq \mathbf{m}\, l + \delta \\ &= \mathbf{m}\, \mathrm{Per}_I(\{u=a\}) + \delta,\end{aligned}$$

where we have used (H1) and (2.11). It suffices to let $\delta \to 0^+$. □

Remark 2.10. The arguments used in the proof of Theorem 2.2 may be easily adapted to generalize the model above to the case where

$$\mathcal{F}_\varepsilon(u) := \begin{cases} \int_\Omega \left(\frac{W(u)}{\varepsilon} + \varepsilon^{2p-1}|\nabla^2 u|^p\right) dx & \text{if } u \in W^{2,p}(\Omega;\mathbb{R}^d), \\ +\infty & \text{if } u \in L^1(\Omega;\mathbb{R}^d) \setminus W^{2,p}(\Omega;\mathbb{R}^d) \end{cases}$$

for $1 < p < +\infty$, with

$$\Gamma(L^1) - \lim_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) = \begin{cases} \mathbf{m}\,\mathrm{Per}_I(\{u = a\}) & \text{if } u \in BV(I; \{a, b\}), \\ +\infty & \text{otherwise,} \end{cases}$$

and now

$$\mathbf{m} := \inf \left\{ \int_{\mathbb{R}} (W(f) + |f''|^p)\, dt : f \in \mathcal{A} \right\}.$$

As usual, the scaling ε^{2p-1} in $\mathcal{F}_\varepsilon$ is the natural one obtained by testing the finiteness of energy with admissible fields u which are a and b in most of the domain, except on a transition layer of width ε. Note that here Proposition 2.3 still applies provided (2.2) is modified to read

$$\|\nabla u\|_{L^q} \le C \left(\|u\|_{L^1}^{1/2} \|\nabla^2 u\|_{L^p}^{1/2} + \|u\|_{L^1} \right)$$

with $2/q = 1 + 1/p$.

Naturally, the next step is to try to understand higher than two perturbations, i.e., how to treat

$$\mathcal{F}_\varepsilon^k(u) := \begin{cases} \int_\Omega \left(\frac{W(u)}{\varepsilon} + \varepsilon^{2k-1} |\nabla^k u|^2 \right) dx & \text{if } u \in W^{k,2}(\Omega; \mathbb{R}^d), \\ \\ +\infty & \text{if } u \in L^1(\Omega; \mathbb{R}^d) \setminus W^{k,2}(\Omega; \mathbb{R}^d), \end{cases}$$

where $k \in \mathbb{N}$. Although the methods involved may stay close to the ones introduced in this paper, this generalization does not seem to follow as immediately as the one above: last, but not least, the corresponding G and H will now require matching of all derivatives up to order $(k-1)$, and a new version of Lemma 2.4 will be in order. This analysis will be carried on in a forthcoming paper.

3. The *N*-dimensional case. Let Ω be an open, bounded, Lipschitz domain in $\mathbb{R}^N$, and consider the functionals

$$\mathcal{F}_\varepsilon(u) := \begin{cases} \int_\Omega \left(\frac{W(u)}{\varepsilon} + \varepsilon^3 |\nabla^2 u|^2 \right) dx & \text{if } u \in W^{2,2}_{\mathrm{loc}}(\Omega; \mathbb{R}^d)\ , \\ \\ +\infty & \text{if } u \in L^1(\Omega; \mathbb{R}^d) \setminus W^{2,2}_{\mathrm{loc}}(\Omega; \mathbb{R}^d) \end{cases}$$

for $\varepsilon > 0$, where W satisfies hypotheses (H1) and (H2). We recall that the constant $\mathbf{m}$ was defined in (2.1).

We start by proving L^1 compactness for energy bounded sequences.

PROPOSITION 3.1. *If* $u_\varepsilon \in W^{2,2}(\Omega; \mathbb{R}^d)$ *satisfy* $\liminf_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u_\varepsilon) < +\infty$*, then there exists a subsequence* $\{u_{\varepsilon_n}\} \subset \{u_\varepsilon\}$ *and* $u \in BV(\Omega; \{a, b\})$ *such that* $u_{\varepsilon_n} \to u$ *in* $L^1(\Omega; \mathbb{R}^d)$.

The proof of this result uses the L^1-slicing compactness criterion introduced by Alberti, Bouchitté, and Seppecher in [2, Theorem 6.6] (see Proposition 3.2 below). Here two sequences $\{u_\varepsilon\}$ and $\{v_\varepsilon\}$ are said to be *δ-close* if $||u_\varepsilon - v_\varepsilon|| < \delta$, $\delta > 0$. When Ω is a rectangle of the form $I \times J$, with I, J open intervals, we write $x \in \Omega$ as $x = (y, z)$ with $y \in I$, $z \in J$. For every function u defined on Ω and every $y \in I$ we denote by u^y the function on J defined by $u^y(z) := u(y, z)$, and for every $z \in J$ we set $u^z(y) := u(y,z)$ for $y \in I$. The functions u^y and u^z are called *one-dimensional slices* of u.

Proposition 3.2. *Assume that the sequence $\{u_\varepsilon\}$ is equi-integrable, and suppose that for every $\delta > 0$ there exist sequences $\{v_\varepsilon\}$ and $\{w_\varepsilon\}$ δ-close to $\{u_\varepsilon\}$, and such that, $\{v^y_\varepsilon\}$ is precompact in $L^1(J;\mathbb{R}^d)$ for a.e. $y \in I$, and $\{w^z_\varepsilon\}$ is precompact in $L^1(I;\mathbb{R}^d)$ for a.e. $z \in J$. Then $\{u_\varepsilon\}$ is precompact in $L^1(\Omega;\mathbb{R}^d)$.*

Remark 3.3. We note that the original statement of Theorem 6.6 in [2] assumes that $\{u_\varepsilon\}$ is bounded in L^∞. However, it is easy to verify that the main tool involved is the use of Fréchet–Kolmogorov theorem for precompactness in L^1, which clearly holds as well when $\{u_\varepsilon\}$ is equi-integrable.

Proof of Proposition 3.1. For simplicity we suppose $N = 2$; the higher dimensional case is treated in an analogous way.

Assume first that Ω is a rectangle of the form $I \times J$, with I, J open intervals.

We denote by $\mathcal{F}^1_\varepsilon(u, A)$ the one-dimensional functional

$$\mathcal{F}^1_\varepsilon(u,A) := \begin{cases} \int_A \left(\frac{W(u)}{\varepsilon} + \varepsilon^3|u''|^2\right) dt & \text{if } u \in W^{2,2}(A;\mathbb{R}^d), \\ +\infty & \text{if } u \in L^1(A;\mathbb{R}^d) \setminus W^{2,2}(A;\mathbb{R}^d), \end{cases}$$

for every open interval A and every $u \in L^1(A;\mathbb{R}^d)$. We recall that if $u \in W^{2,2}(\Omega;\mathbb{R}^d)$, then $u^y \in W^{2,2}(J;\mathbb{R}^d)$ for a.e. $y \in I$ and $u^z \in W^{2,2}(I;\mathbb{R}^d)$ for a.e. $z \in J$, and

$$\frac{\partial^2 u}{\partial z^2}(x) = \frac{d^2 u^y}{dz^2}(z)\ ,\quad \frac{\partial^2 u}{\partial y^2}(x) = \frac{d^2 u^z}{dy^2}(y) \quad \text{for a.e. } x \in \Omega$$

(see [19, section 4.9.2]). Since $|\nabla^2 u|^2 \geq \max\left\{\left|\frac{\partial^2 u}{\partial z^2}\right|^2, \left|\frac{\partial^2 u}{\partial y^2}\right|^2\right\}$, we immediately obtain the following *slicing inequalities*:

$$\mathcal{F}_\varepsilon(u) \geq \int_I \mathcal{F}^1_\varepsilon(u^y, J)\, dy, \qquad \mathcal{F}_\varepsilon(u) \geq \int_J \mathcal{F}^1_\varepsilon(u^z, I)\, dz\,. \tag{3.1}$$

Now consider a family of functions $\{u_\varepsilon\}$ such that $\mathcal{F}_\varepsilon(u_\varepsilon) \leq C < +\infty$. Since $\int_\Omega W(u_\varepsilon)\,dx \leq C\varepsilon$, we have that $W(u_\varepsilon) \to 0$ in L^1, and therefore equi-integrability of $\{u_\varepsilon\}$ follows from (H2). Therefore, fix $\delta > 0$ and let $\delta' \in (0,\delta)$ be such that

$$\mathcal{L}^2(S) \leq \delta'|J| \quad \Longrightarrow \quad \sup_{\varepsilon>0} \int_S (|u_\varepsilon(x)| + |b|)\,dx \leq \delta\,.$$

For $\varepsilon > 0$ we define $v_\varepsilon : \Omega \to \mathbb{R}^d$ by

$$v^y_\varepsilon(z) := \begin{cases} u^y_\varepsilon(z) = u_\varepsilon(y,z) & \text{if } \mathcal{F}^1_\varepsilon(u^y_\varepsilon, J) \leq C/\delta', \\ b & \text{otherwise.} \end{cases}$$

We claim that $v^y_\varepsilon = u^y_\varepsilon$ for all $y \in I$ except at most on a set $Z_\varepsilon \subset I$ of measure smaller than δ'. Indeed, by (3.1) we have

$$C \geq \sup_{\varepsilon>0} \int_I \mathcal{F}^1_\varepsilon(u^y_\varepsilon, J)\,dy, \tag{3.2}$$

and so

$$|Z_\varepsilon| \leq |\{\mathcal{F}^1_\varepsilon(u^y_\varepsilon, J) > C/\delta'\}| \leq \frac{\delta'}{C}\int_I \mathcal{F}^1_\varepsilon(u^y_\varepsilon, J)\,dy \leq \delta',$$

and we have

$$\|u_\varepsilon - v_\varepsilon\|_1 \leq \int_{Z_\varepsilon \times J} |u_\varepsilon(x) - b|\, dx \leq \int_{Z_\varepsilon \times J} (|u_\varepsilon(x)| + |b|)\, dx \leq \delta$$

for every $\varepsilon > 0$, since $\mathcal{L}^2(Z_\varepsilon \times J) \leq \delta' |J|$. Hence the sequence $\{v_\varepsilon\}$ is δ-close to $\{u_\varepsilon\}$. Moreover, for every $y \in I$ there holds $\mathcal{F}^1_\varepsilon(v^y_\varepsilon, J) \leq C/\delta'$, where we have used the fact that $\mathcal{F}^1_\varepsilon(b, J) = 0$, and therefore Proposition 2.7 yields $L^1(J; \mathbb{R}^d)$ precompactness of $\{v^y_\varepsilon\}$. Similarly, we can construct a sequence $\{w_\varepsilon\}$ δ-close to $\{u_\varepsilon\}$ so that $\{w^z_\varepsilon\}$ is precompact in $L^1(I; \mathbb{R}^d)$ for every $z \in J$, and it suffices to now use Proposition 3.2 to conclude that the sequence $\{u_\varepsilon\}$ is precompact in $L^1(\Omega; \mathbb{R}^d)$.

The case where Ω is a general open subset of $\mathbb{R}^N$ is obtained by decomposing Ω into a countable union of closed rectangles with disjoint interiors.

The fact that the limit function u belongs to $BV(\Omega; \{a, b\})$ is showed in the proof of Proposition 3.5. □

THEOREM 3.4. *If $u \in L^1(\Omega; \mathbb{R}^d)$, then*

$$\Gamma(L^1) - \lim_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) = \begin{cases} \mathbf{m}\, \mathrm{Per}_\Omega(\{u = a\}) & \text{if } u \in BV(\Omega; \{a, b\}), \\ +\infty & \text{otherwise.} \end{cases}$$

We divide the proof of this theorem into two propositions concerning, respectively, the $\Gamma(L^1) - \liminf$ and the $\Gamma(L^1) - \limsup$. Although nowadays these arguments may be considered to be quite standard, and we refer the reader to [7, 16], and to [4] for the treatment of second derivatives in the study of the $\Gamma(L^1) - \limsup$; we included here the proofs of Proposition 3.5 and Proposition 3.6 for completeness and for the convenience of the reader.

PROPOSITION 3.5. *Let $u \in L^1(\Omega; \mathbb{R}^d)$. If $\Gamma(L^1) - \liminf_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) < +\infty$, then u belongs to $BV(\Omega; \{a, b\})$ and*

$$\Gamma(L^1) - \liminf_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) \geq \mathbf{m}\, \mathrm{Per}_\Omega(\{u = a\}).$$

Proof. Suppose that $\varepsilon_n \to 0^+$, $u_n \to u$ in $L^1(\Omega; \mathbb{R}^d)$ and $\mathcal{F}_{\varepsilon_n}(u_n)$ converges to $\Gamma(L^1) - \liminf_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) < +\infty$. Fixing an unit vector $\nu \in \mathbb{S}^{N-1}$, possibly passing to a subsequence (not relabelled), we may assume that $u_n|_{L_{y,\nu} \cap \Omega} \to u|_{L_{y,\nu} \cap \Omega}$ in $L^1(L_{y,\nu}, \mathcal{H}^1)$ for almost every line $L_{y,\nu}$ parallel to ν, where $L_{y,\nu} := \{y + s\nu : s \in \mathbb{R}\}$, $y \in \mathbb{R}^N$. By Proposition 2.8, and setting

$$u^{y,\nu}_n(t) := u_n(y + t\nu) \ \text{ for } \ \mathcal{H}^{N-1} \text{ a.e.}\, y \in \nu^\perp,$$

we have

$$\mathbf{m} \frac{|Du^{y,\nu}|(L_{y,\nu} \cap \Omega)}{|b - a|} \leq \liminf_{n \to \infty} \int_{L_{y,\nu} \cap \Omega} \left(\frac{W(u^{y,\nu}_n)}{\varepsilon_n} + \varepsilon^3_n \left| \frac{d^2 u^{y,\nu}_n}{dt^2} \right|^2 \right) dt.$$

Thus, by Fatou's lemma and the slicing properties of BV functions (see [19, section 5.10.2]),

$$\begin{aligned}
\mathbf{m}\,\mathrm{Per}_\Omega(\{u=a\}) &= \mathbf{m}\,\frac{|Du|(\Omega)}{|b-a|}\\
&= \frac{\mathbf{m}}{|b-a|}\int_{\{y\in\nu^\perp\}} |Du^{y,\nu}|(L_{y,\nu}\cap\Omega)\,d\mathcal{H}^{N-1}(y)\\
&\le \int_{\{y\in\nu^\perp\}} \liminf_{n\to\infty}\int_{L_{y,\nu}\cap\Omega}\left(\frac{W(u_n^{y,\nu})}{\varepsilon_n}+\varepsilon_n^3\left|\frac{d^2u_n^{y,\nu}}{dt^2}\right|^2\right)dt\,d\mathcal{H}^{N-1}\\
&\le \liminf_{n\to\infty}\int_{\{y\in\nu^\perp\}}\int_{L_{y,\nu}\cap\Omega}\left(\frac{W(u_n)}{\varepsilon_n}+\varepsilon_n^3|\nabla^2u_n|^2\right)dt\,d\mathcal{H}^{N-1}\\
&= \liminf_{n\to\infty}\int_\Omega\left(\frac{W(u_n)}{\varepsilon_n}+\varepsilon_n^3|\nabla^2u_n|^2\right)dx\\
&= \Gamma(L^1)-\liminf_{\varepsilon\to0^+}\mathcal{F}_\varepsilon(u). \qquad \square
\end{aligned}$$

Proposition 3.6. *For every function $u\in BV(\Omega;\{a,b\})$ we have*

$$\Gamma(L^1)-\limsup_{\varepsilon\to0^+}\mathcal{F}_\varepsilon(u)\le\mathbf{m}\,\mathrm{Per}_\Omega(\{u=a\}).$$

Proof. Let $u\in BV(\Omega;\{a,b\})$, with $u=a\chi_E+b(1-\chi_E)$, and where E is a set of finite perimeter, i.e., $\mathrm{Per}_\Omega(E)=|D\chi_E|(\Omega)<+\infty$. Since E can be approximated by a sequence of smooth sets $E_i=\widetilde{E}_i\cap\Omega$ such that $\widetilde{E}_i$ is a smooth bounded set in $\mathbb{R}^N$, $\chi_{E_i}\to\chi_E$ in $L^1(\Omega)$ and $|D\chi_{E_i}|(\Omega)\to|D\chi_E|(\Omega)$ (see Lemma 4.3 in [6]), in order to study the $\Gamma(L^1)-\limsup$ it suffices to consider a function $u:\Omega\to\mathbb{R}$ such that

$$u(x)=\begin{cases}a & \text{if } x\in E,\\ b & \text{if } x\in\Omega\setminus E,\end{cases}$$

where $E=\widetilde{E}\cap\Omega$ and $\widetilde{E}$ is a compact subset of $\mathbb{R}^N$ with a C^2 boundary. We claim that

$$\Gamma(L^1)-\limsup_{\varepsilon\to0^+}\mathcal{F}_\varepsilon(u)\le\mathbf{m}\,\mathrm{Per}_\Omega(\{u=a\}).$$

Since $M:=\partial\widetilde{E}$ is a C^2 manifold in $\mathbb{R}^N$, there exists $\delta_0>0$ such that for all $\delta<\delta_0$ the points in the tubular neighborhood U_δ of the manifold M admit a unique smooth projection onto M, where $U_\delta:=\{x\in\mathbb{R}^N:\ \mathrm{dist}(x,M)<\delta\}$.

Let $\varepsilon_n\to0^+$, and consider a sequence of functions $v_n\in W^{2,2}_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^d)$ such that

$$v_n(t)=\begin{cases}a & \text{if } t\le-\frac{1}{\sqrt{\varepsilon_n}},\\ b & \text{if } t\ge\frac{1}{\sqrt{\varepsilon_n}},\end{cases}$$

and

$$\lim_{n\to\infty}\int_\mathbb{R}(W(v_n)+|v_n''|^2)\,dt=\mathbf{m}\,.$$

We define the sequence of functions $u_n:\Omega\to\mathbb{R}$

$$u_n(x):=\begin{cases}v_n(\frac{\widetilde{d}_M(x)}{\varepsilon_n}) & \text{if } x\in U_n\cap\Omega,\\ a & \text{if } x\in E\setminus U_n,\\ b & \text{if } x\in\Omega\setminus(E\cup U_n),\end{cases}$$

where $\widetilde{d}_M : \mathbb{R}^N \to \mathbb{R}$ is the signed distance function from the boundary of $\widetilde{E}$, negative inside $\widetilde{E}$, and $U_n := U_{\sqrt{\varepsilon_n}}$. We have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathcal{F}_{\varepsilon_n}(u_n) = \limsup_{n\to\infty} \int_\Omega \left(\frac{W(u_n)}{\varepsilon_n} + \varepsilon_n^3 \left|\nabla^2 u_n\right|^2 \right) dx \\
&= \limsup_{n\to\infty} \left\{ \int_{U_n} \frac{W(v_n(\widetilde{d}_M(x)/\varepsilon_n))}{\varepsilon_n}\, dx + \int_{U_n} e_n^3 \left| v_n'' \nabla \widetilde{d}_M \times \nabla \widetilde{d}_M/\varepsilon_n^2 + v_n' H/\varepsilon_n \right|^2 dx \right\},
\end{aligned}$$

where H is the Hessian matrix of $\widetilde{d}_M$. Change variables via the diffeomorphism $x := F(y,t)$, where $F : M \times (-\delta_0/2, \delta_0/2) \to U_{\delta_0/2}$, $F(y,t) := y + t\nu(y)$, with $\nu(y)$ the normal vector to M at y pointing outside $\widetilde{E}$. We indicate by $J(y,t)$ the Jacobian of this transformation. Then

$$\begin{aligned}
\limsup_{n\to\infty} \mathcal{F}_{\varepsilon_n}(u_n) \le \liminf_{n\to\infty} \Bigg\{ \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \Bigg(& \frac{W(v_n(t/\varepsilon_n))}{\varepsilon_n} \\
& + \varepsilon_n^3 \frac{\left|v_n''(t/\varepsilon_n)\right|^2}{\varepsilon_n^4} \left| \nabla \widetilde{d}_M(F(y,t)) \right|^2 \Bigg) J(y,t)\, dt\, d\mathcal{H}^{N-1}(y) \\
+ \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \varepsilon_n^3 \frac{\left|v_n'(t/\varepsilon_n)\right|^2}{\varepsilon_n^2} & \left|H(F(y,t))\right|^2 J(y,t)\, dt\, d\mathcal{H}^{N-1}(y) \\
+ 2 \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \varepsilon_n^3 \frac{|v_n''(t/\varepsilon_n)||v_n'(t/\varepsilon_n)|}{\varepsilon_n^3} & \left| \nabla \widetilde{d}_M(F(y,t)) \right| |H(F(y,t))| J(y,t) dt d\mathcal{H}^{N-1}(y) \Bigg\},
\end{aligned}$$

which reduces to

$$\begin{aligned}
&\limsup_{n\to\infty} \mathcal{F}_{\varepsilon_n}(u_n) \\
&\le \limsup_{n\to\infty} \Bigg\{ \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \left(\frac{W(v_n(t/\varepsilon_n))}{\varepsilon_n} + \frac{\left|v_n''(t/\varepsilon_n)\right|^2}{\varepsilon_n} \right) J(y,t)\, dt\, d\mathcal{H}^{N-1}(y) \\
&\quad + A \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \varepsilon_n \left|v_n'(t/\varepsilon_n)\right|^2 dt\, d\mathcal{H}^{N-1}(y) \\
&\quad + B \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} |v_n''(t/\varepsilon_n)|\, |v_n'(t/\varepsilon_n)|\, dt\, d\mathcal{H}^{N-1}(y) \Bigg\} \\
&=: \limsup_{n\to\infty} \left\{ I_1^{(n)}(u) + I_2^{(n)}(u) + I_3^{(n)}(u) \right\},
\end{aligned}$$

where we took into account the facts that the gradient of the distance is always equal to one, and that the Jacobian J and the Hessian H of the distance are uniformly bounded. We have

$$\begin{aligned}
I_1^{(n)}(u) &= \int_M \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \left(\frac{W(v_n(t/\varepsilon_n))}{\varepsilon_n} + \frac{\left|v_n''(t/\varepsilon_n)\right|^2}{\varepsilon_n} \right) J(y,t)\, dt\, d\mathcal{H}^{N-1}(y) \\
&= \int_M \int_{-1/\sqrt{\varepsilon_n}}^{1/\sqrt{\varepsilon_n}} \left(W(v_n(s)) + \left|v_n''(s)\right|^2 \right) J(y, s\varepsilon_n)\, ds\, d\mathcal{H}^{N-1}(y) \\
&\le \left(\sup_{y\in M,\, t\in(-\sqrt{\varepsilon_n}, \sqrt{\varepsilon_n})} J(y,t) \right) \int_M \int_{\mathbb{R}} \left(W(v_n(s)) + \left|v_n''(s)\right|^2 \right) ds\, d\mathcal{H}^{N-1}(y)\,,
\end{aligned}$$

and passing to the limit in n as $n \to \infty$, we get

$$\begin{aligned}\limsup_{n\to\infty} I_1^{(n)} &\leq \left(\sup_{y\in M,\, t\in(-\sqrt{\varepsilon_n},\sqrt{\varepsilon_n})} J(y,t)\right) \mathbf{m}\,\mathcal{H}^{N-1}(M)\\ &= \left(\sup_{y\in M,\, t\in(-\sqrt{\varepsilon_n},\sqrt{e_n})} J(y,t)\right) \mathbf{m}\ \mathrm{Per}_\Omega(\{u=a\}).\end{aligned}$$

If we show that the other two integrals $I_2^{(n)}(u)$ and $I_3^{(n)}(u)$ go to zero as $n \to \infty$, then we obtain that

$$\begin{aligned}\Gamma(L^1) - \limsup_{\varepsilon\to 0^+} \mathcal{F}_\varepsilon(u) &\leq \limsup_{n\to\infty} \left(\sup_{y\in M,\, t\in(-\sqrt{\varepsilon_n},\sqrt{\varepsilon_n})} J(y,t)\right) \mathbf{m}\ \mathrm{Per}_\Omega(\{u=a\})\\ &= \mathbf{m}\ \mathrm{Per}_\Omega(\{u=a\}),\end{aligned}$$

where we used the fact that $\{\varepsilon_n\}$ is an arbitrary sequence converging to zero, and that since M is compact, $J(y,t)$ goes uniformly to one as $t \to 0$.

Finally,

$$\begin{aligned}I_2^{(n)} + I_3^{(n)} &\leq C \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \left(\varepsilon_n |v_n'(t/\varepsilon_n)|^2 + |v_n''(t/\varepsilon_n)|\,|v_n'(t/\varepsilon_n)|\right) dt\\ &= C \int_{\mathbb{R}} \left(\varepsilon_n^2 |v_n'(s)|^2 + \varepsilon_n |v_n''(s)|\,|v_n'(s)|\right) ds\\ &\leq C\, \left(\varepsilon_n^2 \|v_n'\|_2^2 + \varepsilon_n \|v_n''\|_2\, \|v_n'\|_2\right)\\ &\leq C\, \left(\varepsilon_n^2 \|v_n'\|_2^2 + \varepsilon_n\, \|v_n'\|_2\right), \qquad (3.3)\end{aligned}$$

where we changed variables, used Hölder inequality and the fact that

$$\|v_n''\|_{L^2}^2 \leq \int_{\mathbb{R}} (W(v_n) + |v_n''|^2)\, dt \to \mathbf{m}\,.$$

Set $w_n(t) := v_n\,(t/\varepsilon_n)$. Then

$$\begin{aligned}&\limsup_{n\to\infty} \int_{-1}^{1} \left(\frac{1}{\varepsilon_n} W(w_n) + \varepsilon_n^3 |w_n''|^2\right) dt\\ &\qquad = \limsup_{n\to\infty} \int_{-\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} \left(\frac{1}{\varepsilon_n} W(w_n) + \varepsilon_n^3 |w_n''|^2\right) dt\\ &\qquad = \lim_{n\to\infty} \int_{\mathbb{R}} (W(v_n) + |v_n''|^2)\, dt,\end{aligned}$$

and therefore by Proposition 2.7 and (2.7)

$$||w_n'||_{L^{4/3}(-1,1)} \leq C\varepsilon_n^{-3/4},$$

where the constant C is independent of n. Also, Hölder inequality yields

$$\begin{aligned}||w_n''||_{L^{4/3}(-1,1)} &= ||w_n''||_{L^{4/3}(-\sqrt{\varepsilon_n},\sqrt{\varepsilon_n})}\\ &\leq C||w_n''||_{L^2(-1,1)}\ \varepsilon_n^{1/6}\\ &\leq C\,\varepsilon_n^{-3/2}\ \varepsilon_n^{1/6}.\end{aligned}$$

Therefore, by the Sobolev embedding theorem

$$||w_n'||_{L^2(-1,1)} \leq C||w_n'||_{W^{1,4/3}(-1,1)} \leq C(\varepsilon_n^{-3/4} + \varepsilon_n^{-3/2}\,\varepsilon_n^{1/6}),$$

and in view of (3.3), we conclude that

$$\begin{aligned}\varepsilon_n^2 \int_R |v_n'|^2\,dt &= \varepsilon_n^3 \int_{\sqrt{\varepsilon_n}}^{\sqrt{\varepsilon_n}} |w_n'|^2\,dt \\ &\leq C\,\varepsilon_n^3\,(\varepsilon_n^{-3/4} + \varepsilon_n^{-3/2}\,\varepsilon_n^{1/6})^2 \\ &\leq C\,(\varepsilon_n^{3/2} + \varepsilon_n^{1/3}),\end{aligned}$$

and it suffices to let $n \to \infty$. □

4. Final remarks. As in the singular perturbation model for phase transitions (see [25]), the interfacial energy appears due to the need to go across an energy barrier in order to remain on the zero set of W. Indeed, if the zero set of W is a smooth, connected set, then the $\Gamma(L^1)$-limit may simply reduce to zero. As an example, consider the case where $\{W = 0\} = \mathbb{S}^{d-1}$. Then

$$\Gamma(L^1) - \lim_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) = \begin{cases} 0 & \text{if } u \in L^1(\Omega; \mathbb{S}^{d-1}), \\ +\infty & \text{otherwise.} \end{cases}$$

To prove this assertion, fix $u \in L^1(\Omega; \mathbb{S}^{d-1})$ and let $\{u_n\}$ be a sequence of smooth functions with compact support, converging to u in $L^1(\Omega; \mathbb{S}^{d-1})$. The existence of such an approximating sequence can be obtained as follows: there exists a point $y \in \mathbb{S}^{d-1}$ such that $u^{-1}(y)$ has zero Lebesgue measure; therefore we may assume with no loss of generality that u does not take such value y. The manifold $\mathbb{S}^{d-1} \setminus \{y\}$ is diffeomorphic to the open unit ball B of $\mathbb{R}^{d-1}$ via some smooth map Φ; hence it is sufficient to approximate the function $\Phi(u)$ in $L^1(\Omega; B)$ with a sequence of smooth functions $\{v_n\}$ with compact support and then to consider the sequence $u_n := \Phi^{-1}(v_n)$.

If now we choose a positive sequence $\varepsilon_n \to 0^+$ such that

$$\int_\Omega \varepsilon_n^3 |\nabla^2 u_n|^2\,dx \leq \frac{1}{n} \qquad \text{for every } n \in \mathbb{N},$$

we get

$$\Gamma(L^1) - \lim_{\varepsilon \to 0^+} \mathcal{F}_\varepsilon(u) \leq \liminf_{n \to \infty} \int_\Omega \left(\frac{W(u_n)}{\varepsilon_n} + \varepsilon_n^3 |\nabla^2 u_n|^2 \right) dx \leq \lim_{n \to \infty} \frac{1}{n} = 0,$$

and this proves the claim.

Finally, we remark that if we could prove that for energy bounded sequences $\{u_\varepsilon\}$, with

$$\sup_{\varepsilon > 0} \int_\Omega \left(\frac{W(u_\varepsilon)}{\varepsilon} + \varepsilon^3 |\nabla^2 u_\varepsilon|^2 \right) dx < +\infty,$$

it follows that

$$\sup_{\varepsilon > 0} \int_\Omega \varepsilon |\nabla u|^2\,dx < +\infty;$$

then most proofs would be greatly simplified, and, in particular, the compactness in L^1 (see Propositions 2.7, 3.1) would follow immediately from the compactness for the singular perturbations model studied in [10, 12, 13, 22, 25, 28, 29, 32, 33].

Acknowledgments. The authors are indebted to Luc Tartar for many enlightening discussions on this problem, and for his continuous interest and advice as the work progressed. Also, the authors would like to thank Luigi Ambrosio and Frédéric Hélein for several interesting and stimulating discussions on the subject of this paper, and Giovanni Alberti for having suggested the use of Proposition 3.2 to obtain compactness of energy bounded sequences in the vectorial case (see Proposition 3.1).

REFERENCES

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

[2] G. Alberti, G. Bouchitté, and P. Seppecher, *Phase transition with the line-tension effect*, Arch. Rational Mech. Anal., 144 (1998), pp. 1–46.

[3] R. Alicandro, A. Braides, and M. S. Gelli, *Free-discontinuity problems generated by singular perturbation*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 1115–1129.

[4] R. Alicandro and M. S. Gelli, *Free-discontinuity problems generated by singular perturbation: The n-dimensional case*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.

[5] L. Ambrosio, C. De Lellis, and C. Mantegazza, *Line energies for gradient vector fields in the plane*, Calc. Var., 9 (1999), pp. 327–355.

[6] L. Ambrosio, I. Fonseca, P. Marcellini, and L. Tartar, *On a volume-constrained variational problem*, Arch. Rational Mech. Anal., 149 (1999), pp. 23–47.

[7] L. Ambrosio, N. Fusco, and D. Pallara, *Special Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, London, 1999.

[8] P. Aviles and Y. Giga, *The distance function and defect energy*, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 923–938.

[9] P. Aviles and Y. Giga, *On lower semicontinuity of a defect energy obtained by a singular limit of the Ginzburg–Landau type energy for gradient fields*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 1–17.

[10] S. Baldo, *Minimal interface criterion for phase transitions in mixtures of Cahn–Hilliard fluids*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 37–65.

[11] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDEs and Continuum Models of Phase Transitions, Springer, Berlin, 1989, pp. 207–215.

[12] A. C. Barroso and I. Fonseca, *Anisotropic singular perturbations—the vectorial case*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 527–571.

[13] G. Bouchitté, *Singular perturbations of variational problems arising from a two-phase transition model*, Appl. Math. Optim., 21 (1990), pp. 289–314.

[14] G. Bouchitté, C. Dubs, and P. Seppecher, *Regular approximation of free-discontinuity problems*, Math. Models Methods Appl. Sci., to appear.

[15] G. Bouchitté, C. Dubs, and P. Seppecher, *Transitions de phases avec un potentiel dégénéré à l'infini, application à l'équilibre de petites gouttes*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 1103–1108.

[16] A. Braides, *Approximation of Free-Discontinuity Problems*, Springer-Verlag, Berlin, 1998.

[17] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Birkhäuser, Boston, 1993.

[18] A. De Simone, R. V. Kohn, S. Müller, and F. Otto, *A compactness result in the gradient theory of phase transitions*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.

[19] L. C. Evans and R. F. Gariepy, *Lectures Notes on Measure Theory and Fine Properties of Functions*, CRC Press, Ann Arbor, 1992.

[20] I. Fonseca, *Phase transitions of elastic solid materials*, Arch. Rational Mech. Anal., 107 (1989), pp. 195–223.

[21] I. Fonseca and L. Tartar, *Second order singular perturbations for nonlinear elastic materials*, in preparation.

[22] I. Fonseca and L. Tartar, *The gradient theory of phase transitions for systems with two potential wells*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 89–102.

[23] E. Gagliardo, *Proprietà di alcune classi di funzioni in più variabili*, Ricerche Mat., 7 (1958), pp. 102–137.

[24] G. Gioia and M. Ortiz, *The morphology and folding patterns of buckling-driven thin-film blisters*, J. Mech. Phys. Solids, 42 (1994), pp. 531–559.

[25] M. E. Gurtin, *On phase transitions with bulk, interfacial, and boundary energy*, Arch. Rational Mech. Anal., 96 (1986), pp. 243–264.

[26] W. Jin, *Singular Perturbations, Fold Energies and Micromagnetics*, Ph.D. thesis, New York University, New York, 1997.

[27] W. Jin and R. V. Kohn, *Singular perturbations and the energy of folds*, J. Nonlinear Sci., to appear.

[28] R. V. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[29] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[30] L. Modica and S. Mortola, *Un esempio di* Γ*-convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[31] L. Nirenberg, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup. Pisa (3), 13 (1959), pp. 115–162.

[32] N. C. Owen, *Nonconvex variational problems with general singular perturbations*, Trans. Amer. Math. Soc., 310 (1988), pp. 393–404.

[33] P. Sternberg, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Rational Mech. Anal., 101 (1988), pp. 209–260.

[34] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Vol. IV, Pitman, Boston, 1979, pp. 136–212.

SIAM J. MATH. ANAL.
Vol. 31, No. 5, pp. 1144–1156

VANISHING SHEAR VISCOSITY IN THE EQUATIONS OF COMPRESSIBLE FLUIDS FOR THE FLOWS WITH THE CYLINDER SYMMETRY[*]

HERMANO FRID[†] AND VLADIMIR SHELUKHIN[‡]

Abstract. We analyze the question of the limit process when the shear viscosity goes to zero for global solutions to the Navier–Stokes equations for compressible heat conductive fluids for the flows which are invariant over cylindrical sheets.

Key words. Navier–Stokes equations, compressible fluids, vanishing shear viscosity

AMS subject classifications. Primary, 35B40, 35B35; Secondary, 35K55

PII. S003614109834394X

1. Introduction.

We study the limit of global solutions to the Navier–Stokes equations of compressible heat conductive fluids as the shear viscosity μ tends to zero. The problem of small viscosity finds many applications, for example, in the boundary layer theory [11]. We restrict ourselves to the flows between two circular coaxial cylinders and assume that the corresponding solutions depend only on the radial variable x, $\Omega = \{x : 0 < a < x < b\}$ and the time variable $t \in (0,T)$. The reduced system of the three-dimensional equations is now of the form [6]

$$\rho_t + (\rho u)_x + \frac{\rho u}{x} = 0, \tag{1.1}$$

$$\rho\left(u_t + uu_x - \frac{v^2}{x}\right) + p_x - (\lambda + 2\mu)\left(u_x + \frac{u}{x}\right)_x = 0, \qquad p = \gamma\rho\theta, \tag{1.2}$$

$$\rho\left(v_t + uv_x + \frac{uv}{x}\right) - \mu\left(v_x + \frac{v}{x}\right)_x = 0, \tag{1.3}$$

$$\rho(w_t + uw_x) - \mu\left(w_{xx} + \frac{w_x}{x}\right) = 0, \tag{1.4}$$

$$c_v\rho(\theta_t + u\theta_x) - \kappa\left(\theta_{xx} + \frac{\theta_x}{x}\right) + p\left(u_x + \frac{u}{x}\right) - \mathcal{Q} = 0, \tag{1.5}$$

$$\mathcal{Q} = \lambda\left(u_x + \frac{u}{x}\right)^2 + \mu\left\{\left(v_x - \frac{v}{x}\right)^2 + w_x^2 + 2u_x^2 + 2\left(\frac{u}{x}\right)^2\right\}. \tag{1.6}$$

[*] Received by the editors August 27, 1998; accepted for publication (in revised form) May 7, 1999; published electronically May 2, 2000.

http://www.siam.org/journals/sima/31-5/34394.html

[†] Instituto de Matemática, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, CEP 21945-970, RJ, Brazil (hermano@impa.br). The research of this author was partially supported by CNPq, proc. 352871/96-2.

[‡] Lavrentyev Institute of Hydrodynamics, Novosibirsk, 630090, Russia (shelukhin@hydro.nsc.ru). The research of the second author was supported by CNPq, proc. 301783/96-9.

Here ρ is the mass density, p is the pressure, θ is the temperature. The velocity vector $\mathbf{v} = (u, v, w)$ is given by the radial, angular, and axial velocities, respectively. We consider fluids obeying the equation of state for a polytropic fluid $p = \gamma\rho\theta$ and the Duhem inequalities $\mu \geq 0$ and $3\lambda + 2\mu \geq 0$ ($\nu = \lambda + 2\mu$) [7]. The constants $\gamma, c_v, \kappa, \lambda$, and μ are considered positive. We put $c_v = 1$; this can always be achieved by a suitable choice of units.

In the domain $Q = (0, T) \times \Omega$, we consider the initial boundary value problem given by (1.1)–(1.6) and

$$\mathbf{v} = 0, \qquad \theta_x = 0 \quad \text{at} \quad \partial\Omega, \tag{1.7}$$

$$(\rho, \mathbf{v}, \theta)|_{t=0} = (\rho_0(x), \mathbf{v}_0(x), \theta_0(x)). \tag{1.8}$$

The results that follow are valid also if the velocity vector satisfies nonhomogeneous boundary conditions corresponding to the symmetry-conserving motions of the bounding cylinders. We discuss the homogeneous case only to simplify the presentation.

Our goal is both to establish the existence and uniqueness of strong global solutions to problem (1.6)–(1.8) and to justify the passage to limit as the shear viscosity goes to zero. In particular, we prove that solutions converge in $L^2(Q)$, as $\mu \downarrow 0$, to a weak solution of (1.1)–(1.6) with $\mu = 0$. The last result can be treated as a new one in the mathematical theory of compressible fluids with $\mu = 0$ and $\lambda > 0$ [10, 12, 13].

Existence theorems for the solutions with the axial symmetry were obtained only for radial flows with $v = w = 0$ [8]. The shear viscosity limit problem was studied earlier for the flows between two parallel plates [4, 14, 15]. Notice that in the case of cylinder symmetry the velocity components influence each other not only through the energy equation but also through the momentum equations. This is the main difference between the system considered in [15] and system (1.1)–(1.6).

It should be emphasized that in sending μ to zero we keep λ fixed and positive at the same time. It enables us to control the derivatives of u to some extent but not of v and w. There are several interesting mathematical issues here. We mention two of them. The first is the strong convergence of v as $\mu \downarrow 0$. One should prove it in order to justify the passage to limit in (1.2). The bounds $\|v\|_{L^\infty(Q)} \leq c$ and $\|\mu x v_x^2\|_{L^1(Q)} \leq c$, which are shown to be uniform in μ, enable us to conclude that $v \to \overline{v}$ weakly-star in $L^\infty(Q)$ and $\mu x v_x^2 \to \chi$ weakly in $\mathcal{M}(Q)$, the space of signed Radon measures on Q, for some function $\overline{v} \in L^\infty(Q)$ and a nonnegative measure χ. Here and in what follows we use the superposed bar to denote a weak limit. Without any new bound for the derivatives of v at hand, we prove the strong convergence $v \to \overline{v}$ in $L^k(Q)$ for any $k \in [1, \infty)$ and the equality $\chi = 0$. The key idea is to verify that the limit equation

$$(x\bar{\rho}\overline{v})_t + (x\bar{\rho}\bar{u}\overline{v})_x + \bar{\rho}\bar{u}\overline{v} = 0 \quad \text{in} \quad D^{'}(Q) \tag{1.9}$$

for the weak limits $\bar{u}$, $\bar{v}$, and $\bar{\rho}$, as $\mu \downarrow 0$, is inherited with the equation

$$(x\bar{\rho}\overline{v}^2)_t + (x\bar{\rho}\bar{u}\overline{v}^2)_x + 2\bar{\rho}\bar{u}\overline{v}^2 = 0 \quad \text{in} \quad D^{'}(Q), \tag{1.10}$$

which can be derived formally from (1.9) by multiplying by $\bar{v}$ and accounting for the limit equation $(x\bar{\rho})_t + (x\bar{\rho}\bar{u})_x = 0$, in $D^{'}(Q)$. To check (1.10), we use the Lagrangian variables associated with the limit flow; this makes our approach different from that of the DiPerna and Lions [2] developed for the linear transport equation. On the other hand, it follows from (1.3) in the limit that

$$(x\bar{\rho}\overline{v^2})_t + (x\bar{\rho}\bar{u}\overline{v^2})_x + 2\bar{\rho}\bar{u}\overline{v^2} + 2\chi = 0 \quad \text{in} \quad D^{'}(Q). \tag{1.11}$$

The comparison of (1.10) and (1.11) results in the equalities $\chi = 0$ and $\overline{v^2} = \bar{v}^2$. The last implies the strong convergence $v \to \overline{v}$ at least in $L^2(Q)$.

The above idea of improvement of the weak convergence $v \to \overline{v}$ to strong goes back to the notion of renormalization introduced into PDE by DiPerna and Lions [2] and developed in [3, 4, 7, 15, 17].

One more mathematical issue is related to the convergence of the dissipation $\mathcal{Q}$ and the rate of pressure work $-p(u_x + \frac{u}{x})$, as $\mu \downarrow 0$. Both these functions are bounded in $L^1(Q)$ uniformly in μ, but we do not obtain any more bound for them. We also do not prove that u_x converges strongly. This is why the passage to limit in the energy equation is not simple. Here the key idea is to study their sum $J \equiv \mathcal{Q} - p(u_x + \frac{u}{x})$, which admits the representation $J = F(u_x + \frac{u}{x}) + J_1$, where

$$(1.12)\; F = -p + (\lambda + 2\mu)\left(u_x + \frac{u}{x}\right), \qquad J_1 = \mu\left(v_x + \frac{v}{x}\right)^2 + \mu w_x^2 - \frac{2\mu}{x}(u^2+v^2)_x.$$

The function F was introduced by Hoff [3] and he called it the *effective viscous flux* since it can be rewritten in an invariant form as $-p + (\lambda + 2\mu)\mathrm{div}\mathbf{v}$. Hoff showed that the variable F is smoother than either of its summands. The effective viscous flux helps here also. We manage to prove that $\overline{J_1} = 0$ and $\overline{F(u_x + u/x)} = F_0(\bar{u}_x + \bar{u}/x)$, where $F_0 = -\gamma\bar{\rho}\bar{\theta} + \lambda(\bar{u}_x + \bar{u}/x)$ is the effective viscous flux of the limit flow. For that we again exploit the technique exposed above.

Let us formulate the main results. Assume that the initial data verify the conditions

$$(1.13)\quad \mathbf{v}_0 \in W_0^{1,2}(\Omega); \quad \rho_0, \theta_0 \in W^{1,2}(\Omega); \quad \|\rho_0^{-1}, \theta_0^{-1}\|_{C(\overline{\Omega})} < \infty; \quad \rho_0 > 0, \theta_0 > 0.$$

From here on we use the notation $\|a, b, \dots\|^2 = \|a\|^2 + \|b\|^2 + \cdots$ for functions $a, b, \dots$ belonging to a functional space equipped with a norm $\|\cdot\|$. For later use we denote by $\|f\|$, $\|f\|_{p,\Omega}$, $\|f\|_{p,Q}$, and $\|f\|_{p,q}$ the norms in $L^2(\Omega)$, $L^p(\Omega)$, $L^p(Q)$, and $L^q(0,T; L^p(\Omega))$, respectively.

THEOREM 1.1. *Under the assumption* (1.13) *there is a unique solution of problem* (1.1)–(1.8) *such that*

$$\mathbf{v}, \theta \in L^\infty(0,T; W^{1,2}(\Omega)) \cap L^2(0,T; W^{2,2}(\Omega)),$$

$$\rho \in L^\infty(0,T; W^{1,2}(\Omega)); \quad \mathbf{v}_t, \theta_t, \rho_t \in L^2(Q), \qquad \rho > 0, \theta > 0.$$

THEOREM 1.2. *There is a sequence $\mu_n \downarrow 0$ such that the corresponding solutions $(\rho, \mathbf{v}, \theta)$ of problem* (1.1)–(1.8) *with $\mu = \mu_n$ converge as follows:*

$$u \xrightarrow{L^s(Q)} \overline{u}; \quad \theta \xrightarrow{L^r(Q)} \overline{\theta}; \quad \rho, v, w \xrightarrow{L^p(Q)} \overline{\rho}, \overline{v}, \overline{w}$$

for any $s \in [1,6)$, $r \in [1,2)$, and $p \in [1,\infty)$. In addition, $\overline{u} \in L^2(0,T; W_0^{1,2}(\Omega))$, $\overline{\theta}_x \in L^q(Q)$, $\overline{\rho} \in BV(Q)$, and the weak convergences

$$u \overset{L^2(0,T;W_0^{1,2}(\Omega))}{\rightharpoonup} \overline{u}, \qquad \theta_x \overset{L^q(Q)}{\rightharpoonup} \overline{\theta_x}, \qquad \nabla_{x,t}\rho \overset{\mathcal{M}(Q)}{\rightharpoonup} \nabla_{x,t}\overline{\rho}$$

hold where $\mathcal{M}(Q)$ is the set of the Radon measures on Q and $q \in [1, 3/2)$. The limit functions solve (1.1)–(1.8) *with $\mu = 0$ in the following sense:*

$$(1.14)\qquad \int_Q \overline{\rho}(\varphi_t + \overline{u}\varphi_x)x\,dx\,dt + \int_\Omega \rho_0\varphi(0,x)\,x\,dx = 0,$$

$$\int_Q \left(\overline{\rho}\overline{u}\psi_t + (\overline{\rho}\overline{u}^2 + \gamma\overline{\rho}\overline{\theta} - \lambda\overline{u}_x - \lambda\frac{\overline{u}}{x})\psi_x\right) x\,dx\,dt \tag{1.15}$$

$$+\int_Q \left(\overline{\rho}(\overline{v}^2 - \overline{u}^2) + \gamma\overline{\rho}\overline{\theta} - \lambda\overline{u}_x - \lambda\frac{\overline{u}}{x}\right)\psi\,dx\,dt + \int_\Omega \rho_0 u_0 \psi(0,x)x\,dx = 0,$$

$$\int_Q \overline{\rho}\overline{v}\Big(\psi_t + \overline{u}\psi_x - \frac{\overline{u}}{x}\psi\Big)x\,dx\,dt + \int_\Omega \rho_0 v_0 \psi(0,x)x\,dx = 0, \tag{1.16}$$

$$\int_Q \overline{\rho}\overline{w}(\psi_t + \overline{u}\psi_x)xdxdt + \int_\Omega \rho_0 w_0 \psi(0,x)x\,dx = 0, \tag{1.17}$$

$$\int_Q \left(\overline{\rho}\overline{\theta}\varphi_t + (\overline{\rho}\overline{\theta}\overline{u} - \kappa\overline{\theta}_x)\varphi_x + \Big(-\gamma\overline{\rho}\overline{\theta}\Big(\overline{u}_x + \frac{\overline{u}}{x}\Big) + \lambda\Big(\overline{u}_x + \frac{\overline{u}}{x}\Big)^2\Big)\varphi\right) x\,dx\,dt \tag{1.18}$$

$$+\int_\Omega \rho_0\theta_0\varphi(0,x)x\,dx = 0.$$

Here, φ and ψ are test functions such that φ, $\psi \in C^1(\overline{Q})$, $\varphi(T,x) = \psi(T,x) = 0$ for any $x \in \Omega$ and $\psi \in W_0^{1,2}(\Omega)$ for any $t \in (0,T)$.

We will sometimes employ the notation $f_r = f_x + f/x$. Observe that the operator $f \to f_r$ has the properties

$$(gh)_r = g_r h + g h_x, \qquad g_{rx} = g_{xr} - \frac{g}{x^2}, \tag{1.19}$$

$$g'(\beta)((\rho\beta)_t + (\rho u\beta)_r) = (\rho g(\beta))_t + (\rho u g(\beta))_r = 0, \tag{1.20}$$

$$\int_\Omega \varphi\psi_r x\,dx = -\int_\Omega \psi\varphi_x x\,dx, \qquad \int_\Omega x\xi_r\,dx = b\xi(b) - a\xi(a) \tag{1.21}$$

for any functions $\beta \in C^1(\overline{Q})$, $g \in C^1(\mathbb{R})$, $\varphi \in D(\Omega)$, $\psi \in D(\Omega)$, and $\xi \in D(\mathbb{R})$.

With this notation, the system (1.1)–(1.8) becomes

$$\rho_t + (\rho u)_r = 0, \tag{1.22}$$

$$(\rho u)_t + (\rho u^2)_r - \frac{\rho v^2}{x} + p_x - \nu u_{rx} = 0, \qquad p = \gamma\rho\theta, \tag{1.23}$$

$$(\rho v)_t + (\rho u v)_r + \frac{\rho u v}{x} - \mu v_{rx} = 0, \tag{1.24}$$

$$(\rho w)_t + (\rho u w)_r - \mu w_{xr} = 0, \tag{1.25}$$

$$(\rho\theta)_t + (\rho u\theta)_r - \kappa\theta_{xr} + p u_r - \mathcal{Q} = 0, \tag{1.26}$$

$$\mathcal{Q} = \nu u_r^2 + \mu v_r^2 + \mu w_x^2 - \frac{2\mu}{x}(u^2 + v^2)_x. \tag{1.27}$$

2. A priori estimates independent of $\boldsymbol{\mu}$. First of all, it follows from (1.22) that

$$\frac{d}{dt}\int_\Omega \rho x\,dx = 0. \tag{2.1}$$

This is our first a priori estimate since we look for solutions such that $\rho > 0$ and $\theta > 0$. Next, by (1.22)–(1.27), (1.19), and (1.20), we may write the equation for the total energy $e = \mathbf{v}^2/2 + \theta$ in the form

$$(\rho e)_t + (\rho u e)_r = \nu(uu_r)_r + \mu(vv_r)_r + \mu(ww_x)_r + \kappa\theta_{xr} - (pu)_r - \frac{2\mu}{x}(u^2+v^2)_x. \tag{2.2}$$

Hence,

$$\frac{d}{dt}\int_\Omega \rho e x\,dx = 0, \tag{2.3}$$

and this is our second estimate.

Another consequence of system (1.22)–(1.27) is the equation

$$\left(\rho\psi(\theta) + \gamma\rho\psi\left(\frac{1}{\rho}\right) + \rho\frac{\mathbf{v}^2}{2}\right)_t + \left(\rho u\left(\frac{\mathbf{v}^2}{2} + \psi(\theta) + \gamma(1+\ln\rho)\right)\right)_r$$

$$= \nu(uu_r)_r + \mu(vv_r)_r + \mu(ww_x)_r - p_x u + \psi^{'}(\theta)(\mathcal{Q} + \kappa\theta_{xr}) - pu_r,$$

where $\psi(s) = s - \ln s - 1$.

Again, we integrate with respect to the measure xdx to obtain

$$\frac{d}{dt}\int_\Omega \rho\left(\psi(\theta) + \gamma\psi\left(\frac{1}{\rho}\right) + \frac{\mathbf{v}^2}{2}\right)x\,dx + \int_\Omega\left(\kappa\frac{\theta_x^2}{\theta^2} + \frac{\mathcal{Q}}{\theta}\right)x\,dx = 0. \tag{2.4}$$

This is the third estimate since $\mathcal{Q} \geq 0$.

Lemma 2.1. *There is a constant c such that $\|\rho\|_{\infty,Q} \leq c$.*

Proof. First we note that this lemma provides the key estimate in the whole development, and the proof goes back to the early paper of Kazhikhov and Shelukhin [5] where considerations were made in the Lagrangian variables.

We write (1.23) in the form

$$(\rho u)_t + \left(\rho u^2 + p - \nu u_r + \int_a^x \frac{\rho}{y}(u^2 - v^2)dy\right)_x = 0.$$

Hence, the function

$$\varphi(t,x) = \int_0^t\left(\nu u_r - \rho u^2 - p - \int_a^x \frac{\rho}{y}(u^2 - v^2)dy\right)dt + \int_a^x \rho_0 u_0 dy$$

satisfies the equalities

$$\varphi_x = \rho u, \quad \varphi_t = \nu u_r - \rho u^2 - p - \int_a^x \frac{\rho}{y}(u^2 - v^2)dy. \tag{2.5}$$

Observe that $\|x\varphi_x\|_1^2 \leq \|x\rho\|_1\|x\rho u^2\|_1$ and the integral $\int_\Omega \varphi x\,dx$ equal to

$$\int_0^t \int_\Omega \left(-\rho u^2 - p - \int_a^x \frac{\rho}{y}(u^2 - v^2)\,dy\right) xdxdt + \int_\Omega x \int_a^x \rho_0 u_0\,dy\,dx$$

is uniformly bounded in $L^\infty(0,T)$ by (2.4). Thus, $\|\varphi\|_{\infty,Q} \leq c$.

Given a function $F(\varphi)$, we compute the material derivative $D_t(\rho F) \equiv (\frac{\partial}{\partial t} + u\frac{\partial}{\partial x})\rho F$. Using (2.5), we have

$$D_t(\rho F) = -\rho F u_r + \rho F' \left(\nu u_r - p - \int_a^x \frac{\rho}{y}(u^2 - v^2)\,dy\right).$$

The choice $F(\varphi) = \exp\frac{\varphi}{\nu}$ results in

$$D_t(\rho F) \leq \rho F \left|\frac{1}{\nu}\int_a^x \frac{\rho}{y}(u^2 - v^2)dy\right| \leq c\rho F.$$

Thus, the lemma is proved.

Lemma 2.2. *There is a constant c such that $\rho \geq c > 0$ and $\|\theta\|_{\infty,1} \leq c$.*

Proof. By (2.5) and with $G = \exp(-\frac{\varphi}{\nu})$, we compute

$$D_t\frac{G}{\rho} = \gamma G\theta + \frac{G}{\nu\rho}\int_a^x \frac{\rho}{y}(u^2 - v^2)dy \leq c\left(\frac{G}{\rho} + \|\theta\|_\infty\right).$$

Hence,

$$\|\frac{1}{\rho}\|_\infty \leq c\left(1 + \int_0^t \|\theta(s)\|_\infty ds\right). \tag{2.6}$$

Now, by the inequality

$$\theta^{\frac{1}{2}}(t,x) - \theta^{\frac{1}{2}}(t,y) \leq \frac{1}{2}\left\|\frac{\theta_x}{\theta}\right\|\|\rho\theta\|_1^{\frac{1}{2}}\left\|\frac{1}{\rho}\right\|_\infty^{\frac{1}{2}},$$

we conclude that

$$\|\theta\|_\infty \leq 2\|x\rho\theta\|_1 + \frac{1}{2a^2}\left\|\frac{\sqrt{x}\theta_x}{\theta}\right\|^2\|x\rho\theta\|_1\left\|\frac{1}{\rho}\right\|_\infty. \tag{2.7}$$

Combining inequalities (2.6) and (2.7), we obtain the validity of the lemma.

Lemma 2.3. *There is a constant c such that $\theta \geq c > 0$.*

Proof. We multiply (1.26) by $-z$, $z = 1/\theta$ to get $\rho D_t z \leq \kappa z_{xr} + \frac{\gamma^2\rho^2}{4\lambda}$. Consequently,

$$(\rho z^N)_t + (\rho u z^N)_r \leq \kappa z^N_{xr} + NA(t)z^N,$$

where $N \in \mathbb{N}$ and the function $A(t)$ is bounded in $L^1(0,T)$ and does not depend on N. Now we integrate the last inequality to arrive at $\|x\rho z^N\|_1 \leq \exp(Nc)\,\|x\rho_0 z_0^N\|_1$. Here, the constant c does not depend on N. Hence, the assertion of the lemma follows.

Lemma 2.4. *There is a constant c such that*

$$\|u_x, \sqrt{\mu}v_x, \sqrt{\mu}w_x\|_{2,Q} \leq c, \quad \|u\|_{\infty,2} \leq c, \quad \|uu_x\|_{3/2,Q} + \|u\|_{6,Q} + \|u\|_{\infty,4} \leq c.$$

Proof. Integrating (1.26), we obtain

$$\int_Q (\nu u_r^2 + \mu v_r^2 + \mu w_x^2) x \, dx \, dt \le \frac{\lambda}{2} \int_Q u_r^2 x \, dx \, dt + \frac{\gamma^2}{2\lambda} \int_Q \rho^2 \theta^2 x \, dx \, dt + c.$$

On the other hand,

$$\|\theta\|_{2,Q} \le \|\theta\|_{1,\infty} \|\theta\|_{\infty,1}. \tag{2.8}$$

Hence, the first estimate of the lemma is proved. The second one is now a consequence since u is bounded in $L^\infty(0,T;L^2(\Omega))$. The third estimate follows due to the inequalities $\|uu_x\|_{3/2,Q} \le \|u_x\|_{2,Q}\|u\|_{6,Q}$ and

$$\|u\|_{6,Q}^6 \le \|u\|_{2,\infty}^2 \|u\|_{\infty,4}^4, \quad \|u\|_{\infty,4}^2 \le 2\|u\|_{2,\infty}\|u_x\|_{2,Q}.$$

Lemma 2.5. *There is a constant c such that $\|v,w\|_{\infty,Q} \le c$.*

Proof. Multiplication of (1.22) by $2Nw^{2N-1}$ results in the inequality

$$(\rho w^{2N})_t + (\rho u w^{2N})_r \le \mu (w^{2N})_{xr}.$$

Hence, $\|x\rho w^{2N}\|_1 \le \|x\rho_0 w_0^{2N}\|_1$. Now, we raise to the power $\frac{1}{2N}$ and let $N \to \infty$ to see that $\|w\|_\infty \le \|w_0\|_\infty$.

Similarly,

$$\int_\Omega \rho v^{2N} x dx \le \int_\Omega \rho_0 v_0^{2N} x dx - 2N \int_0^t \int_\Omega \rho u v^{2N} dx ds.$$

By the Gronwall lemma, $\|\rho v^{2N} x\|_1 \le c \exp\left(cN \int_0^t \|u\|_\infty ds\right)$. Since the constant c does not depend on N, the estimate $|v| \le c$ follows.

Lemma 2.6. *There is a constant c such that $\|\rho_x\|_{1,\infty} \le c$ and $\|\rho_t\|_{1,Q} \le c$.*

Proof. We set $\beta = u - \nu(1/\rho)_x$ and find, by (1.22) and (1.23), that

$$(\rho\beta)_t + (\rho u \beta)_x + \frac{\rho}{x}(u^2 - v^2) = -p_x.$$

Multiplying this equality by sgnβ, we deduce

$$(\rho|\beta|)_t + (\rho u|\beta|)_x \le c(1 + \theta|\beta| + \theta|u| + u^2 + v^2 + |\theta_x|).$$

On the other hand, due to (2.4) and (2.8),

$$\|\theta_x\|_{1,Q} \le c. \tag{2.9}$$

Hence, $\|\rho\beta\|_1 \le c$ uniformly in time. Now the first conclusion of the lemma follows directly and the second one holds due to (1.22).

Lemma 2.7. *Given a number $q \in [1, 3/2)$, there is a constant c such that $\|\theta_x\|_{q,Q} \le c$.*

Proof. By above estimates, we may treat (1.26) as a linear parabolic one

$$\rho(\theta_t + u\theta_x) = \kappa\Big(\theta_{xx} + \frac{\theta_x}{x} + f\Big),$$

where the function $f(t,x)$ is bounded in $L^1(Q)$ uniformly in μ. Given a function $F(s) \in C^1(\mathbb{R})$, we have

$$\frac{d}{dt}\int_\Omega \rho F(\theta) x \, dx + \kappa \int_\Omega \theta_x^2 F^{''}(\theta) x \, dx = \int_\Omega f F^{'}(\theta) x \, dx.$$

Let us choose

$$F(\theta) = \frac{1}{1-\delta} + \frac{1+\theta}{\delta} - \frac{(1+\theta)^{1-\delta}}{\delta(1-\delta)},$$

with $\delta \in (0,1)$. We deduce that

$$\kappa \left\| \frac{\theta_x^2}{(1+\theta)^{1+\delta}} \right\|_{1,Q} \leq c_\delta(\|f\|_{1,Q} + \|\theta_0\|_1).$$

As shown in [15] (see also [9]), this inequality implies the bound formulated in the lemma.

We complete this section by proving Theorem 1.1. We argue like in [1, 14], so what follows is a brief scheme. It should be noted that from now until the end of this section the viscosity μ is considered fixed.

Proof of Theorem 1.1. First we prove by the Faedo–Galerkin method that the solution described in Theorem 1.1 exists locally. To this end, we use the finite-dimensional spaces

$$X_n^1 = \operatorname{span}\{\sin E_j(x), j = 1,\dots,n\}, \quad X_n^2 = \operatorname{span}\{\cos E_j(x), j = 1,\dots,n\},$$

$E_j(x) = \frac{j\pi(x-a)}{b-a}$, with the corresponding orthogonal projections $P_n^i : L^2(\Omega) \to X_n^i$, $i \in \{1,2\}$. We look for functions $\mathbf{v}^n \in [X_n^1]^3$, $\theta^n \in X_n^2$, and ρ^n, satisfying

$$(x\rho^n)_t + (x\rho^n u^n)_x = 0, \quad \rho^n|_{t=0} = \rho_0(x),$$

$$P_n^1(xM_n^j(t)) = 0, \quad P_n^2(xM_n^4(t)) = 0, \quad j \in \{1,2,3\},$$

$$\mathbf{v}^n(0) = P_n^1(\mathbf{v}_0), \quad \theta^n(0) = P_n^2(\theta_0),$$

where $M_n^j(t)$, $j \in \{1,2,3,4\}$, are the left-hand sides of (1.2)–(1.5), respectively, with the functions $\mathbf{v}$, θ, and ρ substituted by $\mathbf{v}^n$, θ^n, and ρ^n.

One can prove by means of the standard fixed point arguments (see [14] for details) that approximations exist on some interval $[0,T_n]$, $T_n \leq T$, with $\mathbf{v}^n \in C^1([0,T_n];[X_n^1]^3)$, $\theta^n \in C^1([0,T_n];[X_n^2])$, and ρ^n,ρ_x^n,$\rho_t^n \in L^\infty(0,T_n;L^2(\Omega))$. The next step is to verify that all the approximations are bounded uniformly in n on a same time interval $[0,T_*]$ in the norms defining the solution class in Theorem 1.1. This enables us to conclude that at least some subsequence of approximations converges on $[0,T_*]$ to a solution given by Theorem 1.1.

To prove the global existence, it remains to extend the local solution to the entire interval $[0,T]$. One can do it by deriving global estimates dependent on μ keeping the same line of arguments as in [14]. This completes the proof of global existence. As for the uniqueness, the method proposed in [14] is also applicable here.

3. Vanishing shear viscosity.

We send μ to zero and consider the problem of μ-dependence of solutions $\mathbf{s}_\mu = (\rho,\mathbf{v},\theta)$ of (1.1)–(1.8). It is implicit that the functions $\rho,\mathbf{v}$, and θ depend on μ.

Let us denote

$$C^1(\overline{Q})^T = \{\varphi \in C^1(\overline{Q}) : \varphi(T,x) = 0, x \in \Omega\},$$

$$C^1(\overline{Q})^\Pi = \{\varphi \in C^1(\overline{Q})^T : \varphi|_{\partial\Omega} = 0\}.$$

We need the following general lemma which is a chain rule formula for distribution derivatives. It can be proved by standard functional analysis considerations (see, e.g., [15]).

Lemma 3.1. *Let Ω be a bounded domain in $\mathbb{R}^N$ and $Q = (0,T) \times \Omega$. Assume $A \in L^2(0,T; W_0^{1,2}(\Omega))$, $A_0 \in L^2(\Omega)$, and $\mathbf{B}, C \in L^2(Q)$. If the equality*

$$\int_Q (A\psi_t + \mathbf{B} \cdot \nabla\psi + C\psi)dxdt + \int_\Omega A_0\psi(0,x)dx = 0$$

holds for any $\psi \in C^1(\overline{Q})^\Pi$, then the equality

$$\int_Q \left(\frac{A^2}{2}\psi_t + \mathbf{B} \cdot \nabla(A\psi) + CA\psi\right) dxdt + \int_\Omega \frac{A_0^2}{2}\psi(0,x)dx = 0$$

holds for any $\psi \in C^1(\overline{Q})^T$.

Proof of Theorem 1.2. The proof is divided into several steps. When we speak of a convergence $\mathbf{s}_\mu \to \mathbf{s}$ we will always mean that there is a sequence $\mu_n \downarrow 0$ such that $\mathbf{s}_{\mu_n} \to \mathbf{s}$. Let the vector $\overline{\mathbf{s}} = (\overline{\rho}, \overline{\mathbf{v}}, \overline{\theta})$ stand for the weak limit in $L^2(Q)$ of the sequence $\mathbf{s}_\mu$. Clearly, this limit exists by the above estimates.

Step 1. The fact that $\rho \to \overline{\rho}$ in $L^q(Q)$, $1 \le q < \infty$, follows from the uniform boundedness of ρ in $W^{1,1}(Q) \cap L^\infty(Q)$ and the Sobolev imbedding theorem.

Let us consider the sequence ρu, $\mu \downarrow 0$. By the uniform estimates, it follows from (1.23) that the sequence $(\rho u)_t$, $\mu \downarrow 0$, is also uniformly bounded in $L^2(0,T; W^{-1,2}(\Omega))$. Next, due to the inequality $\|\rho_x u\|_{1,Q} \le \|u\|_{\infty,1}\|\rho_x\|_{1,\infty}$, the sequence $(\rho u)_x$, $\mu \downarrow 0$, is bounded in $L^1(Q)$. Thus, by the Aubin–Simon theorem [16], $\rho u \to \overline{\rho u}$ in $L^2(0,T; L^1(\Omega))$. Now the inequality

$$|u^\mu - u^\nu| \le \frac{|\rho^\mu u^\mu - \rho^\nu u^\nu|}{\rho^\mu} + \left|\frac{1}{\rho^\mu} - \frac{1}{\rho^\nu}\right| |\rho^\nu u^\nu|$$

implies that $u \to \overline{u}$ in $L^1(Q)$. Since u is uniformly bounded in $L^6(Q)$ this convergence is valid in $L^q(Q)$, $q \in [1,6)$. Moreover, $\overline{u} \in L^6(Q)$.

In a similar manner we study the sequence θ, $\mu \downarrow 0$. By the above estimates and from (1.26) it follows that the sequence $(\rho\theta)_t$, $\mu \downarrow 0$, is bounded at least in $L^1(0,T; W^{-2,2}(\Omega) + L^1(\Omega))$. Again, by the inequality $\|\rho_x\theta\|_{1,Q} \le \|\theta\|_{\infty,1}\|\rho_x\|_{1,\infty}$, we conclude that the sequence $(\rho\theta)_x$, $\mu \downarrow 0$, is bounded in $L^1(Q)$. Thus, $\theta \to \overline{\theta}$ in $L^q(Q)$ for any $q \in [1, 3/2)$.

Step 2. Here we prove a strong convergence of v and w as $\mu \downarrow 0$. With the poor control of derivatives, we argue in a manner quite different from that in Step 1. First, consider the sequence w, $\mu \downarrow 0$. We start from the observation that $\overline{\rho}, \overline{u}$, and $\overline{w}$ satisfy (1.14) and (1.17) with test functions $\varphi,\psi \in W^{1,2}(Q)^T$, the closure of $C^1(\overline{Q})^T$ in $W^{1,2}(Q)$. Indeed, the set $C^1(\overline{Q})^T$ of test functions φ for equality (1.14) can be substituted by $W^{1,2}(Q)^T$ by the continuity argument. As for equality (1.17), we first, by the same argument, substitute the set $C^1(\xi_\delta(x) = \min\{\delta, \text{dist}(x, \partial\Omega)\}$. Clearly, $\psi\xi_\delta/\delta \in W^{1,2}(Q)^\Pi$ for any $\psi \in W^{1,2}(Q)^T$. Now, to justify the extension, one needs only to prove that

$$\lim_{\delta\to 0} \frac{1}{\delta} \int_Q \overline{\rho}\overline{u}\overline{w}\psi\xi_{\delta x} x dxdt = 0.$$

But this equality holds since $\overline{u} \in L^2(0,T;W_0^{1,2}(\Omega))$.

Next, we pass to the Lagrangian coordinates (t,y) by the formulas

$$y(t,x) = a + \int_a^x \overline{\rho}(t,s)s\,ds, \qquad y_x = x\overline{\rho}, \qquad y_t = -x\overline{\rho}\overline{u}. \tag{3.1}$$

Without loss of generality, we may assume that $\int_\Omega x\rho_0 dx = b-a$. Thus, with the change of variables, the functions $\overline{\mathbf{s}}(t,\cdot)$ and $\overline{\mathbf{s}}(t,y)$ are related again to the domains Ω and Q. Now, in the new coordinates, (1.17) for $\overline{w}(t,x)$ with the test set $W^{1,2}(Q)^T$ reads

$$\int_Q \overline{w}(t,y)\Psi_t(t,y)dydt + \int_\Omega w_0(y)\Psi(0,y)dy = 0. \tag{3.2}$$

This integral law holds for all $\Psi \in C^1(\overline{Q})^T$. Really, given $\Psi \in C^1(\overline{Q})^T$, the function $\psi(t,x) = \Psi(t,y(t,x))$ belongs to $W^{1,2}(Q)^T$ in view of (3.1). Clearly, the change of variables (3.1) transforms $\psi(t,x)$ into $\Psi \in C^1(\overline{Q})^T$.

Equation (3.2) implies that $\overline{w}(t,y) = w_0(y)$ almost everywhere (a.e.) on $(0,T)$, hence $\overline{w}^2(t,y) = w_0^2(y)$ a.e. on $(0,T)$ and

$$\int_Q \overline{w}^2(t,y)\Psi_t(t,y)dydt + \int_\Omega w_0^2(y)\Psi(0,y)dy = 0 \tag{3.3}$$

for all $\Psi(t,y) \in C^1(\overline{Q})^T$. By continuity argument, (3.3) holds for all $\Psi(t,y)$ such that

$$\Psi, \Psi_t \in L^1(Q), \qquad \Psi|_{t=T} = 0, \qquad \Psi \geq 0. \tag{3.4}$$

Next, multiplying (1.25) by $2xw\psi$, $\psi \in C^1(\overline{Q})^T$, integrating over Q, and sending μ to zero, we obtain

$$J_1(\psi) \equiv \int_Q \overline{\rho}\overline{w^2}(\psi_t + \overline{u}\psi_x)xdxdt + \int_\Omega \rho_0 w_0^2 \psi(0,x)xdx = 2\langle \overline{\mu x w_x^2}, \psi\rangle. \tag{3.5}$$

Here, $\overline{\mu x w_x^2}$ is a nonnegative Radon measure on Q, a weak limit of $\mu x w_x^2$ in the space of signed Radon measures on Q with finite mass. It is a simple consequence of (3.5) that

$$J_1(\psi) \geq 0 \tag{3.6}$$

for all $\psi \in W^{1,2}(Q)_+^T$, the subindex "+" denoting nonnegativity.

By switching to the Lagrangian coordinates, (3.6) reads

$$\int_Q \overline{w^2}(t,y)\Psi_t(t,y)dydt + \int_\Omega w_0^2(y)\Psi(0,y)dy \geq 0 \tag{3.7}$$

for all $\Psi(t,y) \in C^1(\overline{Q})_+^T$. It is true because any function $\Psi(t,y)$ from $C^1(\overline{Q})_+^T$ is transformed by (3.1) into the function $\psi(t,x) = \Psi(t,y(t,x))$ belonging to $W^{1,2}(Q)_+^T$. By continuity, the set (3.4) fits (3.7) as a test set as well.

Comparing (3.3) and (3.7) on the test set (3.4), we find that $\overline{w^2}(t,y) \leq \overline{w}^2(t,y)$ a.e. in Q. On the other hand, by convexity argument, $\overline{w^2}(t,y) \geq \overline{w}^2(t,y)$ a.e. in Q. Hence, $\overline{w^2}(t,x) = \overline{w}^2(t,x)$ a.e. in Q. This implies that $w \to \overline{w}$ in $L^2(Q)$. Since the sequence $w, \mu \downarrow 0$, is bounded in $L^\infty(Q)$ the last convergence holds also in $L^q(Q)$, $q \in [1,\infty)$.

We treat the sequence v, $\mu \downarrow 0$, in the same manner. Clearly, (1.16) holds with the test set $W^{1,2}(Q)^T$. The switching to the Lagrangian coordinates transforms (1.16) into

$$\int_Q \bigl(\overline{v}(t,y)\Psi_t(t,y) - \frac{\overline{v}(t,y)\bar{u}(t,y)}{x(t,y)}\Psi(t,y)\bigr)\,dy\,dt + \int_\Omega v_0(y)\Psi(0,y)\,dy = 0 \tag{3.8}$$

for all $\Psi \in C^1(\overline{Q})^T$. By Lemma 2.4, one can prove that the set

$$\Psi \in L^{\frac{6}{5}}(Q), \qquad \Psi_t \in L^1(Q), \qquad \Psi|_{t=T} = 0, \qquad \Psi \geq 0 \tag{3.9}$$

fits (3.8) as a test set.

Given $\eta(t,y) \in C^1(\overline{Q})^T$, we choose $\Psi = \eta e^U$, where $U = \int_0^t \overline{u}(s,y)/x(s,y)\,ds$. This choice is possible since $\overline{u} \in L^2(0,T;W_0^{1,2}(\Omega))$.

Denoting $V = \overline{v}(t,y)e^U$, we see that

$$\int_Q V\eta_t\,dy\,dt + \int_\Omega v_0(y)\eta(0,y)\,dy = 0$$

for all $\eta \in C^1(\overline{Q})^T$. Hence, $V(t,y) = v_0(y)$ a.e. on $(0,T)$ and we arrive at the representation formula $\overline{v}(t,y) = v_0(y)e^{-U(t,y)}$. Clearly, $V^2(t,y) = v_0^2(y)$ a.e. on $(0,T)$, i.e.,

$$\int_Q \left(\overline{v}^2(t,y)\Psi_t(t,y) - \frac{2\overline{v}^2(t,y)\bar{u}(t,y)}{x(t,y)}\Psi(t,y)\right)dy\,dt + \int_\Omega v_0^2(y)\Psi(0,y)\,dy = 0 \tag{3.10}$$

for all $\Psi \in C^1(\overline{Q})^T$. Again, we can extend the test set $C^1(\overline{Q})^T$ to (3.9).

Let us multiply (1.24) by $2xv\psi$, integrate over Q, and send μ to zero. As a result we obtain

$$\int_Q \left\{\overline{\rho}\,\overline{v^2}(\psi_t + \overline{u}\psi_x) - \frac{2}{x}\overline{\rho}\,\overline{v^2}\,\overline{u}\psi\right\} x\,dx\,dt + \int_\Omega \rho_0 v_0^2\psi(0,x)x\,dx = 2\langle\overline{\mu x v_x^2},\psi\rangle \geq 0 \tag{3.11}$$

for all $\psi \in C^1(\overline{Q})_+^T$. Here, $\overline{\mu x v_x^2}$ is a nonnegative Radon measure, whose existence is guaranteed by Lemma 2.4. Switching to the Lagrangian coordinates, we transform (3.11) into

$$\int_Q \left(\overline{v^2}\Psi_t - \frac{2}{x}\overline{v^2}\,\overline{u}\Psi\right)dydt + \int_\Omega v_0^2\Psi(0,y)dy \geq 0. \tag{3.12}$$

By the arguments above, the set (3.9) fits (3.12) as a test set. The comparison of (3.10) and (3.12) gives $\overline{v^2}(t,y) = \overline{v}^2(t,y)$. Hence, $v(t,x) \to \overline{v}(t,x)$ in $L^2(Q)$ as $\mu \to 0$. Clearly, this convergence is also valid in $L^q(Q)$, $q \in [1,\infty)$. Now it follows from (3.5) and (3.11) that

$$\langle\overline{\mu x w_x^2},\psi\rangle = 0, \qquad \langle\overline{\mu x v_x^2},\psi\rangle = 0$$

for all $\psi \in C^1(\overline{Q})^T$.

As another consequence of the above strong convergences we have that the functions $\overline{\rho}$, $\overline{\mathbf{v}}$, and $\overline{\theta}$ satisfy (1.15).

Step 3. It remains to prove (1.18). The difficulty lies on the weak limit equalities

$$\overline{-xp\left(u_x + \frac{u}{x}\right)} = -x\gamma\overline{\rho}\overline{\theta}\left(\bar{u}_x + \frac{\bar{u}}{x}\right), \qquad \overline{(\lambda+2\mu)x\left(u_x + \frac{u}{x}\right)^2} = \lambda x\left(\bar{u}_x + \frac{\bar{u}}{x}\right)^2.$$

We do not prove each of them, but rather their sum

$$\overline{xF(u_x+u/x)} = xF_0(\bar{u}_x+\bar{u}/x), \qquad F_0 = -\gamma\bar{\rho}\bar{\theta}+\lambda(\bar{u}_x+\bar{u}/x), \tag{3.13}$$

an equality of measures restricted to act on functions ψ from $C^1(\overline{Q})^T$. Here the effective viscous flux F is defined in (1.12) and F_0 is the effective fiscous flux of the limit flow. We prove (3.13) directly by applying Lemma 3.1. Actually, there is a deeper, more interesting point here: we exploit the fact discovered by Hoff [3] that the variable $-p+(\lambda+2\mu)(u_x+u/x)$ behaves better than either of its summands.

We rewrite (1.15) in the Lagrangian coordinates

$$\int_Q \left(\overline{u}\Psi_t + \left(\bar{\rho}x\Psi_y+\frac{\Psi}{x}\right)\left(\gamma\bar{\theta}-\lambda x\overline{u}_y-\frac{\lambda\overline{u}}{x\bar{\rho}}\right)+\frac{\Psi}{x}\overline{v}^2\right)dydt \tag{3.14}$$

$$+\int_\Omega u_0(y)\Psi(0,y)dy = 0.$$

As above, one can show that (3.14) holds for all $\Psi(t,y)\in C^1(\overline{Q})^\Pi$. By Lemma 3.1,

$$\int_Q\left(\frac{\overline{u}^2}{2}\Psi_t + x\left(\gamma\bar{\rho}\bar{\theta}-\lambda x\bar{\rho}\overline{u}_y-\frac{\lambda\overline{u}}{x}\right)(\overline{u}\Psi)_y\right)dydt \tag{3.15}$$

$$+\int_Q\left(\bar{\rho}\overline{v}^2-\bar{\rho}\overline{u}^2+\gamma\bar{\rho}\bar{\theta}-\lambda x\bar{\rho}\overline{u}_y-\frac{\lambda\overline{u}}{x}\right)\frac{\Psi\overline{u}}{x\bar{\rho}}dydt+\int_\Omega\frac{u_0^2}{2}(y)\Psi(0,y)dy = 0$$

for all $\Psi(t,y)\in C^1(\overline{Q})^T$. Clearly, (3.15) holds for the test set

$$\Psi\in L^\infty(Q), \qquad \Psi_t\in L^{\frac{3}{2}}(Q), \qquad \Psi_y\in L^3(Q), \qquad \Psi|_{t=T}=0. \tag{3.16}$$

Now we rewrite (3.15) in the Eulerian coordinates

$$\int_Q\left(\frac{x\bar{\rho}\overline{u}^2}{2}\psi_t+\psi_x\left(\frac{x\overline{u}^3\bar{\rho}}{2}-\lambda\overline{u}^2+\gamma x\overline{u}\bar{\rho}\bar{\theta}-\lambda x\overline{u}\overline{u}_x\right)\right)dxdt+\frac{1}{2}\int_\Omega x\rho_0u_0^2\psi(0,x)dx \tag{3.17}$$

$$+\int_Q\psi\left(x\overline{u}_x\left(\gamma\bar{\rho}\bar{\theta}-\lambda\overline{u}_x-\frac{\lambda\overline{u}}{x}\right)+\bar{\rho}\overline{u}\overline{v}^2-\bar{\rho}\overline{u}^3+\gamma\bar{\rho}\bar{\theta}\overline{u}-\lambda\overline{u}_x\overline{u}-\frac{\lambda\overline{u}^2}{x}\right)dxdt = 0.$$

This equality is valid for all $\psi\in C^1(\overline{Q})^T$ since the transformation $\psi(t,x)\to\Psi(t,y)$ maps $C^1(\overline{Q})^T$ into set (3.16).

Now we multiply (1.23) by $xu\psi$, $\psi\in C^1(\overline{Q})^T$ and send μ to zero. It results in

$$J_2-\int_Q x\psi(\gamma\bar{\rho}\bar{\theta}\bar{u}_x-\lambda\bar{u}_x^2)dxdt = \langle\overline{-xpu_x+(\lambda+2\mu)xu_x^2},\psi\rangle, \tag{3.18}$$

where J_2 is the left-hand side of equality (3.17). Clearly, equalities (3.18) and (3.13) are equivalent. This completes the proof of Theorem 1.2.

REFERENCES

[1] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov, *Boundary Value Problems in Mechanics of Nonhomogeneous Fluids*, North-Holland, Amsterdam, 1990.

[2] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[3] D. Hoff, *Global solutions of the Navier–Stokes equations for multi-dimensional compressible flow with discontinuous initial data*, J. Differential Equations, 120 (1995), pp. 215–254.

[4] A. V. Kazhikhov and V. V. Shelukhin, *The verification compactness method*, in Current Problems in Modern Mathematics, Vol. 2, Novosibirsk, 1996, pp. 51–60.

[5] A. V. Kazhikhov and V. V. Shelukhin, *Unique global solution in time of initial boundary value problems for one-dimensional equations of a viscous gas*, Prikl. Mat. Meh., 41 (1977), pp. 273–283.

[6] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, 2nd ed., Pergamon Press, Oxford, 1987.

[7] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Vol.* 1: *Incompressible Models*, Clarendon Press, Oxford, 1996.

[8] V. B. Nikolaev, *On solvability of mixed problems for the one-dimensional viscous gas equations of the axisymmetrical motion*, Dinamika Sploshn. Sredy, 44 (1980), pp. 83–92 (in Russian).

[9] J. M. Rakotoson, *A compactness lemma for quasilinear problems: Application to parabolic equations*, J. Funct. Anal., 106 (1992), pp. 358–374.

[10] L. Rosenhead et al., *A discussion on the first and the second viscosities of fluids*, Proc. Roy. Soc. London Ser. A, 226 (1954), pp. 1–69.

[11] H. Schlichting, *Boundary Layer Theory*, 7th ed., McGraw-Hill, New York, 1979.

[12] P. Secchi, *Existence theorems for compressible viscous fluids having zero shear viscosity*, Preprint 38050, Dipartamento di Matematica, Libera Universita di Trento, 38050 POVO (TN) Italy, 1982.

[13] J. Serrin, *Mathematical Principles of Classical Fluids Mechanics*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1959.

[14] V. V. Shelukhin, *A shear flow problem for the compressible Navier–Stokes equations*, Internat. J. Non-Linear Mech., 33 (1998), pp. 247–257.

[15] V. V. Shelukhin, *The limit of zero shear viscosity for compressible fluids*, Arch. Rational Mech. Anal., 143 (1998), pp. 357–374.

[16] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 546 (1987), pp. 65–96.

[17] V. A. Weigant and A. V. Kazhikhov, *On global existence of the two-dimensional Navier–Stokes equations of viscous compressible fluid*, Sibirsk. Mat. Zh., 36 (1995), pp. 1283–1316 (in Russian).

SIAM J. MATH. ANAL.
Vol. 31, No. 5, pp. 1157–1174

ASYMPTOTICS OF THE FAST-DIFFUSION EQUATION WITH CRITICAL EXPONENT*

VICTOR A. GALAKTIONOV[†], LAMBERTUS A. PELETIER[‡], AND JUAN L. VAZQUEZ[§]

Dedicated to the memory of Stanislav N. Kruzhkov

Abstract. We study the large-time behavior of the solutions of the initial-value problem for the nonlinear diffusion equation

$$u_t = \nabla \cdot (u^{-\sigma} \nabla u) \quad \text{in } \mathbf{R}^n \times \mathbf{R}_+ \tag{ND}$$

in dimensions $n \geq 3$ with nonnegative initial data $u(x,0) \in L^1(\mathbf{R}^n)$ when the exponent takes on the *critical value* $\sigma = 2/n$. This represents a borderline case in the study of the problem and offers marked qualitative and technical differences with the neighboring cases $\sigma \approx 2/n$, $\sigma \neq 2/n$. In particular, it marks the transition between two completely different asymptotic behavior types. It is known that solutions exist globally in time and conserve the L^1-norm for this problem. We prove that they decay exponentially in time with a complicated law:

$$\log \|u(\cdot,t)\|_\infty \sim -\kappa M^{-2/(n-2)} t^{n/(n-2)} \quad \text{as} \quad t \to \infty,$$

where $M = \int u(x,0)dx$ is the conserved mass and the constant $\kappa > 0$ depends only on the dimension n. This strongly differs from the comparatively simple self-similar asymptotics of the case $\sigma < 2/n$.

The description is split into an *inner* and an *outer* region, conveniently matched at a *transition layer*. The analysis of the outer region can be done independently and the behavior is governed by a first-order conservation law which acts as the reduced asymptotic equation. The uniqueness theory for first-order conservation laws is one of the great contributions of S. N. Kruzhkov to mathematics. The behavior in the inner parabolic region is then studied by means of a semiconvexity argument which makes it possible to translate into this region the precise behavior from the outer region.

Key words. fast-diffusion equation, Cauchy problem, asymptotic behavior, critical exponent, matched expansions, singular perturbations

AMS subject classifications. 35K55, 35K65

PII. S0036141097328452

1. Introduction.

This paper is devoted to describing the large-time behavior of nonnegative, finite-mass solutions of the Cauchy problem for the *fast-diffusion equation*

$$u_t = \nabla \cdot (u^{-\sigma} \nabla u) \tag{1.1}$$

posed in $Q = \mathbf{R}^n \times \mathbf{R}_+$ with space dimension $n \geq 3$ and *critical exponent*

$$\sigma = 2/n. \tag{1.2}$$

This problem has been much studied for the different values of the diffusivity exponent $\sigma \in \mathbf{R}$, and the appropriate behavior has been rigorously established in a large number of situations. It follows from these studies that the asymptotic behavior of such

[*] Received by the editors October 9, 1997; accepted for publication (in revised form) May 17, 1999; published electronically May 2, 2000. This work was partially supported by TMR project FMRX-CT98-0201 of the EU.

http://www.siam.org/journals/sima/31-5/32845.html

[†] School of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK.

[‡] Mathematical Institute, Leiden University, 2300 RA Leiden, The Netherlands (peletier@math.leidenuniv.nl).

[§] Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

equations depends strongly on the class of initial data $u(x,0)=u_0(x)$; the condition of finite mass

$$\|u_0\|_1=\int_{\mathbf{R}^n} u_0(x)dx<\infty,$$

also termed L^1-data, is the appropriate way of selecting the class of all "small" solutions with similar asymptotics. The case $\sigma=2/n$ represents a critical or borderline case for the class of L^1-solutions and has not been rigorously analyzed. It is our aim to fill this gap. A previous contribution in that direction is due to King, who performed in [Ki] a very detailed *formal* analysis of (1.1) in the so-called fast-diffusion range, i.e., $\sigma>0$. As we shall show, the behavior of the problem with critical exponent has a higher level of complexity, which is due to the fact that it represents the transition from one type of self-similar asymptotic behavior to a different type. Indeed, it is well known that for $\sigma\in(-\infty,2/n)$, the problem has a global-in-time solution which exhibits an asymptotic behavior with self-similarity of the first kind, i.e., determined by dimensional analysis, while for $\sigma\in(2/n,1)$ solutions exist only for a finite time and the extinction behavior is self-similar of the second kind. This will be explained below in greater detail. It is to be noted that in the critical case $\sigma=2/n$, the solution u does not evolve as $t\to\infty$ toward a single global self-similar solution, so a two-region analysis is needed.

Let us now give a more detailed description of the contents of the paper. As mentioned above, we consider sufficiently smooth initial data

$$u(x,0)=u_0(x)\in L^1(\mathbf{R}^n),\quad u_0(x)\geq 0. \tag{1.3}$$

A unique classical solution exists for problem (1.1)–(1.3), and conservation of mass holds [BC]:

$$\int u(x,t)\,dx=\int u_0(x)\,dx=M\in(0,\infty),\quad t>0. \tag{1.4}$$

The solution is positive for all times $t>0$. Our results can be summarized as follows. We establish the decay rate

$$\log\|u(\cdot,t)\|_\infty=-\kappa(n)\,\|u_0\|_1^{-2/(n-2)}t^{n/(n-2)}(1+o(1))\quad\text{as}\quad t\to\infty, \tag{1.5}$$

where $\kappa(n)$ is given by

$$\kappa=n(n-2)(2n\omega_n)^{2/(n-2)}$$

and ω_n denotes the volume of the unit ball in $\mathbf{R}^n$. Moreover, (1.5) gives in first approximation the asymptotic behavior of $\log(u)$ in the whole *inner region*, which is the ball of radius

$$R(t)=\exp\left\{\kappa_0\|u_0\|_1^{-2/(n-2)}t^{n/(n-2)}\right\},\quad \kappa_0=\kappa/n; \tag{1.6}$$

cf. Theorem 3.1. In other words, the profile of $\log(u)$ in the inner region becomes *flat* in first approximation. It has to be added that the solutions become asymptotically radially symmetric as $t\to\infty$. On the other hand, the analysis of the *outer region* $\{|x|>R(t)\}$ performed in section 2 gives a behavior of the form

$$\log u(r,t)\sim\frac{n}{2}\{\log(n-2)+\log t-2\log r-\log\log r\}. \tag{1.7}$$

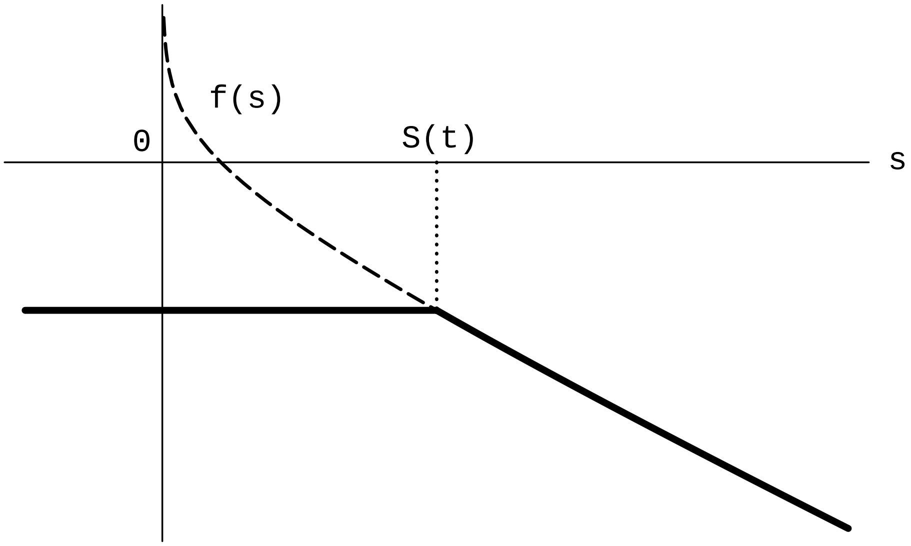


FIG. 1.1. *Asymptotic profile of* $\log(ut^{-n/2})$ *(the bold line) versus* $s = \log r$*,* $f(s) = -ns - \frac{n}{2}\log s + \frac{n}{2}\log(n-2)$*,* $S(t) = \kappa_0 M^{-2/(n-2)} t^{n/(n-2)}$.

Thus, in logarithmic scale, the profile of $u(x,t)$ has a broken shape, as sketched in Figure 1.1.

This behavior was predicted and formally analyzed in [Ki] under conditions of radial symmetry and is rigorously established below. The multiple-region structure is what makes the description of the asymptotic behavior different and more involved than for other diffusivity exponents. The present study and the proofs adapt to that structure. Thus, the outer region can be analyzed independently. After the change of variables

$$v(s,\phi,t) = r^n u(r,\phi,t), \quad s = \log r, \tag{1.8}$$

where we use standard spherical coordinates $x = (r,\phi) \in \mathbf{R}^{\mathrm{n}}$, we find that v satisfies a nonlinear convection-diffusion equation (see (2.2) and (2.24) below), where the convection part is asymptotically dominant. This is first analyzed in the radially symmetric case to which the general problem is then reduced. For radial functions, the asymptotic structure is most easily studied on a rescaled version of v, θ given by

$$\theta(\xi,\tau) = t^{\alpha} v(\xi\, t^{\alpha}, t), \quad \tau = \log t, \quad \alpha = \frac{n}{n-2}, \tag{1.9}$$

whose evolution is governed by a viscous perturbation of a nonlinear first-order conservation law (2.9). An entropy inequality plays a crucial role in the analysis; cf. Proposition 2.1. Together with the stability theorem of [GV1] it allows us to show that the viscous term is asymptotically negligible and to establish the asymptotic profile, which we call an N-wave following the usual terminology in the literature on conservation laws; see Figure 1.2.

We see that the outer region ends at a transition layer which is located at distance $R(t)$, obtained as the position of the shock of the asymptotic wave. The precise convergence result is stated in Theorem 2.3. The results obtained in the radial case

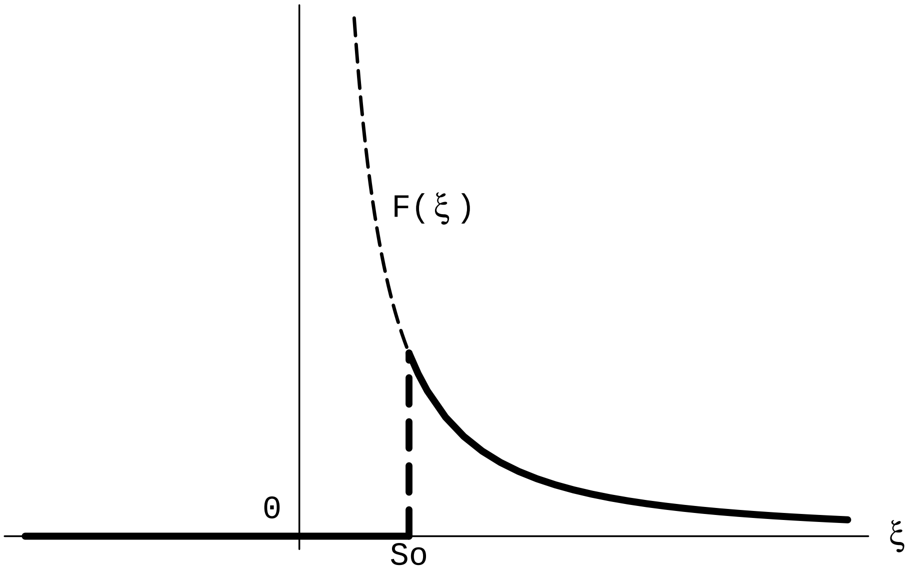


FIG. 1.2. *$v\,t^{n/(n-2)}$ (the bold line) versus $\xi = (\log r)\,t^{-n/(n-2)}$ for $t \gg 1$, $F(\xi) = (\xi/(n-2))^{-n/2}$, $s_0 = \kappa_0 M^{-2/(n-2)}$.*

are extended to general initial data thanks to the property of asymptotic radial symmetry, plus an approximation step; cf. Theorem 2.5. Summing up, the outer region is asymptotically of hyperbolic type and completely determines the whole asymptotic process, even if (1.1) is a purely diffusive equation.

The second step is to describe the behavior of the *inner* region which in terms of v is asymptotically trivial in relative size according to the previous result. However, it cannot be trivial in terms of u since the solutions are monotonically decreasing as functions of r. We thus revert to the original u variable and show that the large-time behavior is controlled by the second-order parabolic operator thanks to a semiconvexity result plus matching with the outer expansion near the transition surface $|x| = R(t)$. This allows us to get the function on the right-hand side of (1.5) as the first term in the asymptotic approximation for $\log(u)$ in the whole inner region.

Let us briefly recall the asymptotic results for $\sigma \neq 2/n$ for the sake of comparison. For $\sigma = 0$ we obtain the classical heat equation, $u_t = \Delta u$, and nonnegative, finite-mass solutions converge asymptotically toward the Gaussian kernel. The next well-studied case is the range $-\infty < \sigma < 0$, where the equation is usually written as $u_t = (1/m)\Delta(u^m)$ with $m = 1-\sigma > 1$ and is known as the porous medium equation. Any solution in our class converges to one of the self-similar *source-type* solutions first described in [ZK] and [B], precisely the one with the same mass; cf. [FK]. Accordingly, the rate of decay takes the form

$$\|u(t)\|_\infty = C(n,\sigma)M^{2k/n}t^{-k}(1+o(1)), \quad t \to \infty, \tag{1.10}$$

with $k = [(2/n) - \sigma]^{-1} > 0$, which for $\sigma = 0$ gives the well-known exponent $n/2$ of the linear case. This analysis can be extended to the subcritical fast-diffusion case, $0 < \sigma < 2/n$: the source-type solutions given by the Zel'dovich–Kompaneetz–Barenblatt formula still exist (cf. [LL]), the mass of the solutions is still conserved in time, and finite-mass solutions converge as $t \to \infty$ to the source-type solution with the same mass. The analysis breaks down as $\sigma \to 2/n$. No source-type solutions exist in this critical case $\sigma = 2/n$ as demonstrated by [BF]. We also see that the exponent $-k$ diverges, thus indicating the change of behavior and suggesting an exponential decay rate.

In the supercritical range $\sigma > 2/n$ there are no solutions with finite mass in dimensions $n = 1$ or 2 (cf. [V]), hence our restriction of the dimension. Solutions with finite mass exist if $n \geq 3$ and $2/n < \sigma < 1$. Mass is not conserved and the solutions undergo extinction in finite time [BC], so that for some $T = T(u_0) < \infty$ there holds $u(x, T) \equiv 0$. In this case (1.1) admits a unique self-similar solution of the second kind [Ki, PZ], which is proved to be asymptotically stable as $t \to T^-$ [GP]. This implies a decay rate of the form

$$\|u(t)\|_\infty = C(n, \sigma)(T - t)^\gamma(1 + o(1)), \quad t \to T^-, \tag{1.11}$$

where $T > 0$ is the extinction time, a function of σ, n, and the initial data, and $\gamma = \gamma(\sigma, n) > 0$ is the anomalous exponent. We point out that in all cases a single analysis gives a uniform asymptotic approximation of the solution and it has self-similar form.

2. Analysis of the outer region. We start our analysis of the behavior of the solutions in the outer region under the additional condition of radial symmetry on the initial data, $u_0 = u_0(r)$, $r = |x|$, so that $u = u(r, t)$ for all $t > 0$. We also impose a number of other conditions whose occurrence will be justified by the properties of the solutions of the equation. Under these assumptions we establish in this section the hyperbolic behavior in the outer region in four steps.

2.1. Radial setting: Change of variables. We consider a finite-mass, radially symmetric solution of the Cauchy problem (1.1)–(1.3). It is natural for such asymptotic problems to assume that $u_0(r) \to 0$ as $r \to \infty$ monotonically. Indeed, after a displacement of the origin of time we may suppose without loss of generality that $u_0(r)$ is decreasing due to the property of eventual monotonicity of such solutions, a proof of which we add as an appendix for the convenience of the reader. It also follows that $u_r \leq 0$ in Q by the maximum principle. For simplicity of the analysis we also impose the following monotonicity assumption on the data v_0:

$$v_0(s) \equiv r^n u_0(r) \quad \text{has a single maximum.} \tag{M}$$

This condition will be true again after shifting the origin of time for solutions with compactly supported initial data and we defer the proof to the appendix. By a suitable approximation in L^1, the results are then extended to general data; see below.

As in [Ki], we begin by performing the change of variables

$$v(s, t) = r^n u(r, t), \quad s = \log r. \tag{2.1}$$

In this way we get the following one-dimensional quasi-linear heat equation with a nonlinear convection term:

$$v_t = (v^{-2/n} v_s)_s - n(v^{(n-2)/n})_s. \tag{2.2}$$

It follows from (M) that by the strong maximum principle applied to (2.2) differentiated with respect to s, we have that for every $t > 0$, the function $v(s, t)$ has a single maximum at a point $s = s_m(t)$ so that $v_s > 0$ for $s < s_m(t)$ and $v_s < 0$ for $s > s_m(t)$.

It turns out that in the outer region the *first-order convection* term is precisely the one controlling the asymptotic behavior in the first approximation.

2.2. Entropy inequality and L^∞-bound. We must show that the first-order term on the right-hand side of (2.2) is dominant. Indeed, the situation of a diffusion-convection equation whose asymptotic behavior is convective was studied in [EVZ1] in the simpler model $v_t = v_{xx} - (v^q)_x$ in the exponent range $1 < q < 2$. See also [EVZ2] for the application to several dimensions. There the convective control manifested itself in the form of an *entropy inequality.* A similar result holds for (2.2). Before stating and proving it and deriving its consequences, let us remark that the asymptotic degeneracy of parabolic equations is a typical feature of finite-time extinction and blowup; cf. [GV2].

PROPOSITION 2.1. *For every smooth solution of* (2.2) *we have*

$$(v^{-2/n})_s \le \frac{1}{(n-2)t}. \tag{2.3}$$

Proof. Let us write (2.2) in terms of $w = v^{-2/n}$:

$$w_t = ww_{ss} - \frac{n}{2}(w_s)^2 - (n-2)ww_s.$$

Take now $z = w_s$. It satisfies

$$z_t = wz_{ss} + [(1-n)z - (n-2)w]z_s - (n-2)z^2.$$

Since this equation admits the explicit solution $Z(t) = 1/(n-2)t$ with unbounded initial data, $Z(+0) = +\infty$, we obtain (2.3) by the maximum principle. □

Remark. This estimate is the cornerstone on which the introduction of the dynamical systems analysis, section 2.3, is based. It can also be written as

$$v_s \ge -\frac{n}{2(n-2)t}\, v^{(n+2)/n}.$$

It is exact if we neglect the diffusion term, as will happen later in the asymptotic limit. Indeed, there is a family of self-similar entropy solutions of the convective equation $v_t = -n(v^{(n-2)/n})_s$ of the form

$$V(s,t) = \begin{cases} \left[\dfrac{s}{(n-2)t}\right]^{-n/2} & \text{for } s \ge s_0 t^{n/(n-2)}, \\ 0 & \text{for } s < s_0 t^{n/(n-2)}, \end{cases} \tag{2.4}$$

where $s_0 > 0$ is a free constant. These discontinuous solutions are called N-waves in the literature; cf. [S]. The free constant s_0 can easily be determined from (2.4) as a function of the preserved mass M of the solution

$$s_0 = \kappa_0 M^{-2/(n-2)}, \quad \kappa_0 = \kappa/n. \tag{2.5}$$

As a consequence, we will show that the N-waves provide the asymptotic profiles and estimate (2.3) will also be optimal for the solutions of the whole equation (2.2).

Before proceeding further we still need some other standard facts. One of them is the conservation of mass for (2.2), i.e., for every solution and every $t > 0$ we have

$$M = \int_{\mathbf{R}^n} u(x,t)\,dx = n\omega_n \int_0^\infty r^{n-1}u(r,t)\,dr = n\omega_n \int_{-\infty}^\infty v(s,t)\,ds. \tag{2.6}$$

On the other hand, the maximum principle implies that v is bounded, and in particular that $v(s,t) \leq \|v(\cdot,0)\|_\infty$. Moreover, all nonnegative and bounded solutions are actually positive and C^∞-smooth.

The entropy inequality and the conservation of mass immediately give an important a priori L^∞-estimate which controls the actual size of the solutions for large times.

Corollary 2.2. *For every solution $v(s,t)$ with mass $M > 0$ we have*

$$v(s,t) \leq CM^{n/(n-2)}t^{-n/(n-2)}. \tag{2.7}$$

Proof. This estimate is a straightforward consequence of the entropy inequality (2.3) and the mass-conservation equation (2.6). See [EVZ1, Lemma 1.2]. □

2.3. Rescaled equation and hypotheses of the stability theorem. We proceed now with the large-time analysis which is based on the dynamical systems approach of [GV1]. The first step in such an analysis is to perform a rescaling which makes the orbits of our evolution problem compact with nontrivial limits. Based on estimate (2.7), we introduce the new variables

$$\theta(\xi,\tau) = t^\alpha v(\xi\, t^\alpha, t), \quad \tau = \log t. \tag{2.8}$$

Here and below the scaling exponent has the value $\alpha = n/(n-2)$. We have the following equation for θ:

$$\theta_\tau = \mathbf{B}(\tau,\theta) \equiv \mathbf{A}(\theta) + \mathbf{e}^{-\alpha\tau}\mathbf{C}(\theta), \tag{2.9a}$$

where

$$\mathbf{A}(\theta) = -\mathbf{n}(\theta^{(\mathrm{n}-2)/\mathrm{n}})_\xi + \alpha(\xi\theta)_\xi \qquad \text{and} \qquad \mathbf{C}(\theta) = (\theta^{-2/\mathrm{n}}\theta_\xi)_\xi. \tag{2.9b}$$

The general approach of [GV1] adapts to our case as follows: we consider (2.9) as an asymptotically small perturbation of the purely convective equation

$$\theta_\tau = \mathbf{A}(\theta), \tag{2.10}$$

which is called the *limit* or *reduced* equation for (2.9). Both equations are viewed as abstract evolution equations posed in a Banach space, in this case $X = L^1(\mathbf{R}^n)$, so that a solution is viewed as a curve $u(\tau) : (0,\infty) \mapsto L^1(\mathbf{R}^n)$. In the case of (2.9) we consider the class $\mathcal{S}$ of solutions obtained by formulas (2.1) and (2.8) from the class of solutions of problem (1.1)–(1.3) stated in the introduction. In the case of the first-order equation (2.10), which is equivalent to

$$v_t = -n(v^{(n-2)/n})_s, \tag{2.11}$$

it is well known that the proper concept of solution is Kruzhkov's entropy solution [Kr], and that the autonomous operator $\mathbf{A}$ generates a semigroup of contractions in $L^1(\mathbf{R})$. We can restrict our consideration to the class $\mathcal{T}$ of nonnegative entropy solutions with fixed finite mass $M > 0$.

In this situation if the three conditions of compactness, consistency, and stability are satisfied we have the following result.

Theorem S (see [GV1]). *The ω-limit sets for the solutions $\theta(\tau) \in \mathcal{S}$ of* (2.9) *are contained in the global ω-limit set Ω^* of* (2.10). *Consequently, the orbits approach Ω^* uniformly as $\tau \to \infty$ in the $L^1(\mathbf{R}^n)$-norm.*

We recall that the ω-limit set of a solution $\theta \in \mathcal{S}$ of (2.9) is defined as

$$\omega(\theta_0) = \{v \in X : \exists\ \tau_j \to \infty \quad \text{such that} \quad \theta(\tau_j) \to v \text{ in } \ X\}. \tag{2.12}$$

The global ω-limit set Ω^* of (2.10) is defined as the closure of the set of all $v \in X$ which can be obtained as limits

$$v = \lim_{t_j \to \infty} \theta(t_j),$$

where $\theta \in \mathcal{T}$ is any solution of (2.10) and $\{t_j\}$ is any sequence which goes to infinity. In fact, only the solutions of (2.11) which can be obtained as limits of solutions of (2.9) need to be considered, and then the set Ω^* receives the name *reduced* ω-limit set of (2.11).

The three conditions to be satisfied are the following.

(H1) *Compactness of the orbits.* We must have a class $\mathcal{S}$ of weak solutions $\theta \in C([0,\infty) : X)$ of (2.9) defined for all $\tau > 0$ with values in $X = L^1(\mathbf{R}^n)$. The orbits $\{\theta(\tau) : \tau > 0\}$ must be relatively compact. Moreover, if we define the shifted orbits

$$\theta^t(\tau) = \theta(\tau + t), \quad t, \tau > 0, \tag{2.13}$$

then the sets $\{\theta^t\}_{t>0}$ must be relatively compact in $L^\infty(0,\infty : X)$.

The verification of this condition is not difficult. In particular, the boundedness of the orbits follows from the agreement between estimate (2.7) and the rescaling (2.8). Compactness comes from (2.3) and standard regularity.

(H2) *Consistency.* Given any solution $\theta \in \mathcal{S}$ and a sequence $t_j \to \infty$ such that θ^{t_j} converges to a function u in $L^\infty(0,\infty : X)$, then $u(\tau)$ is a solution of (2.10) in the class $\mathcal{T}$.

The fact that weak solutions of (2.9) give in the limit weak solutions of (2.10) is immediate. The limits are entropy solutions because of Proposition 2.1.

(H3) *Stability for the limit equation.* The set Ω^* is nonvoid, compact, and uniformly Lyapunov stable.

The Lyapunov stability of the limit set for the first-order equation (2.10), or equivalently (2.11), comes from the fact that we are dealing with a semigroup of contractions in $L^1(\mathbf{R}^n)$. On the other hand, it is known that the ω-limit set for such conservation laws consists of N-wave profiles satisfying the stationary autonomous equation; cf. [LP]. The best-known case in the literature is the equation $v_t = (v^m)_s$ with $m > 1$, where the N-wave has compact support in space. A convergence analysis in that case can be seen in [EVZ1]. The fact that the exponent $(n-2)/n$ in (2.11) is less than 1 implies that the N-wave has unbounded support with an *infinite tail* (2.4). In terms of the variables ξ and θ it reads

$$F(\xi) = \begin{cases} \left(\dfrac{\xi}{n-2}\right)^{-n/2} & \text{for } \xi \geq s_0, \\ 0 & \text{for } \xi < s_0. \end{cases} \tag{2.14}$$

In view of the conservation of mass, (2.6), the final profile is uniquely determined by the constant s_0 given in (2.5). Therefore, the reduced ω-limit set of all orbits with the given total mass of the autonomous equation (2.10) is

$$\Omega_* = \{F = F_*(\xi) \quad \text{with} \quad s_0 = \kappa_0 M^{-2/(n-2)}\}, \tag{2.15}$$

and it is uniformly stable in the L^1-metric. This completes (H3).

The survey paper [GV3] contains a general discussion of the application of these ideas to the study of evolution problems.

2.4. Outer behavior. Thanks to Theorem S we conclude that the ω-limits of the orbits of (2.9) are an N-wave as above and $\omega(\theta_0) \subseteq \Omega_*$. In view of (2.15) this yields that $\omega(\theta_0) = \{F_*(\xi)\}$, i.e., as $\tau \to \infty$,

$$\theta(\cdot,\tau) \to F_*(\cdot) \quad \text{in } L^1(\mathbf{R}). \tag{2.16}$$

In fact, as a straightforward consequence we get a stronger convergence.

THEOREM 2.3. *As $t \to \infty$ we have*

$$v(s,t) = \left[\frac{s}{(n-2)t}\right]^{-n/2} (1 + o(1)) \tag{2.17}$$

uniformly on the sets $\{s \geq (s_0 + \varepsilon)\, t^{n/(n-2)}\}$, where $\varepsilon > 0$ may be arbitrarily small, while

$$v(s,t) = o(t^{-n/(n-2)}) \tag{2.18}$$

uniformly on $\{s \leq (s_0 - \varepsilon)\, t^{n/(n-2)}\}$.

Proof. It follows from the monotonicity assumption (M) and (2.16) that the maximum $\xi_m(\tau)$ of the rescaled function $\theta(\xi,\tau)$ satisfies $\xi_m(\tau) \to s_0$ as $\tau \to \infty$. Therefore, there exists a $\tau_1 \gg 1$ such that

$$\theta_\xi(\xi,\tau) < 0 \qquad \text{for all} \qquad \xi > s_0 + \varepsilon/4, \quad \tau > \tau_1, \tag{2.19}$$

and

$$\theta_\xi(\xi,\tau) > 0 \qquad \text{for all} \qquad \xi < s_0 - \varepsilon/4, \quad \tau > \tau_1.$$

In addition, the entropy inequality (2.3) yields the lower bound

$$\theta_\xi(\xi,\tau) \geq -\frac{n}{2(n-2)}\theta^{(n+2)/n}. \tag{2.20}$$

Thus, in the outer region, for $\xi > s_0 + \varepsilon/4$, we have an upper as well as a lower bound for θ_ξ on compact subsets. This allows us to strengthen the convergence of $\theta(\cdot,\tau)$ to F_* in L^1 to uniform convergence in compact intervals. Translated to the variables s and t, this results in (2.17).

In the inner region, the L^1-convergence can be strengthened to uniform convergence to zero on compact intervals, thanks to the monotonicity of $\theta(\xi,\tau)$ with respect to ξ, and (2.18) follows by monotonicity. □

We thus conclude that convergence (2.16) is uniform on compact subsets in ξ bounded away from the point $\xi = s_0$. In terms of the original variables we get the following *outer-region expansion:*

$$u^{2/n}(r,t) = \frac{(n-2)t}{r^2 \log r}(1 + o(1)) \quad \text{as} \quad t \to \infty \tag{2.21}$$

if $r \geq \exp\{(\kappa_0 + \varepsilon) M^{-2/(n-2)}\, t^{n/(n-2)}\}$. From this outer expansion we obtain an estimate of the decay rate of the form

$$\log u(r,t) \leq -\kappa M^{-2/(n-2)}\, t^{n/(n-2)} (1 + o(1)). \tag{2.22}$$

We will show in section 3 that this estimate is correct uniformly in $\mathbf{R}^n$. Moreover, in the logarithmic scale the solution becomes flat in the inner region and takes the self-similar form given in the right-hand side of (2.22).

2.5. The nonradial case. We now generalize the previous asymptotic behavior to nonradial solutions. A basic ingredient of the proof is the principle of asymptotic symmetry that says that solutions with compactly supported initial data become almost radially symmetric for large enough times. Such results are well known [GNN] and apply to the fast-diffusion equation, as we will see. In this case we define the new variable v by means of standard spherical coordinates $x = (r, \phi)$, $\phi \in S^{n-1}$, and put $s = \log r$ and

$$v(s, \phi, t) = r^n u(r, \phi, t), \tag{2.23}$$

which satisfies the evolution equation

$$v_t = (v^{-2/n} v_s)_s + \frac{n}{n-2} L(v^{(n-2)/n}) - n(v^{(n-2)/n})_s, \tag{2.24}$$

where L is the Laplace–Beltrami operator (cf. (2.2)). With this definition the result of Theorem 2.3 is still true. The passage to general initial data is done by approximation in a second stage.

Compactly supported data. We consider problem (1.1)–(1.2) with continuous and nonnegative initial data u_0 supported in the ball of radius $a > 0$, $B_a(0)$. Let us fix the mass of the solution $M = \int u_0(x)\, dx > 0$. In a first step we use a classical argument based on reflection, due to Aleksandrov and Serrin, which proves the following result.

LEMMA 2.4. *Under the above conditions the solution of the initial-value problem satisfies*

$$u(x, t) \geq u(y, t) \tag{2.25}$$

for every $t > 0$ and every pair of points $x, y \in \mathbf{R}^n$ such that $|y| \geq |x| + a$.

A similar result is used in [CVW] for the porous-medium equation, $\sigma < 0$. As a consequence of this fact, if we consider the radial functions

$$\underline{u}(r, t) = \inf_{|x|=r} u(x, t), \quad \overline{u}(r, t) = \sup_{|x|=r} u(x, t),$$

we will have for $|x| = r > 0$

$$\underline{u}(r, t) \leq u(x, t) \leq \overline{u}(r, t)$$

and also

$$\underline{u}(r, t) \geq \overline{u}(r + a, t).$$

The next step consists of proving that these two radial functions, $\underline{u}$ and $\overline{u}$, which are a lower and upper bound for $u(\cdot, t)$, respectively, have a mass very similar to $u(\cdot, t)$ for large t. In fact, let $\varepsilon > 0$ be small and let us take $T > 0$ large enough such that the solution with initial data $\overline{u}_0(r)$ has mass less than ε inside the ball of radius $R = 2a/\varepsilon$, which is easily seen to be true thanks to the estimates of the previous section. We call

$$f(x) = u(x, T), \quad \underline{f}(r) = \underline{u}(r, T), \quad \overline{f}(r) = \overline{u}(r, T).$$

Then $\int \underline{f}(r)dx \leq M \leq \int \overline{f}(r)dx$. But we also have an estimate in the other direction. Indeed, for some $\lambda = 1+\varepsilon$, $\varepsilon > 0$ very small, we have

$$\underline{f}(r) \geq \overline{f}(\lambda r)$$

if $\lambda r \geq r + a$, i.e., $r \geq A = a/\varepsilon$. Then

$$\int_{|x|\geq A} \overline{f}(x)\,dx \geq \int_{|x|\geq A} \overline{f}(\lambda x)\,dx = \frac{1}{\lambda^n}\int_{|y|\geq \lambda A} \overline{f}(y)\,dy.$$

On the other hand, by the assumption on T the mass of $\overline{f}$ and $\underline{f}$ inside the ball $B_R(0)$ is less than ε. It follows that

$$\int_{\mathbf{R}^n} \overline{f}(x)\,dx \geq (1-n\varepsilon)\int_{\mathbf{R}^n} \overline{f}(x)\,dx - \varepsilon.$$

We easily conclude that

$$M - c\varepsilon \leq \int \underline{f}(x)\,dx \leq \int \overline{f}(x)\,dx \leq M + c\varepsilon, \quad c = n+1.$$

The final step consists of fixing as initial time a time T as above and starting the radial evolution with initial data $\overline{f}$ and $\underline{f}$, for which the asymptotic result of Theorem 2.3 is true. Then we observe that u is in between and that $\varepsilon \to 0$ as $t \to \infty$.

General case. When we assume that u_0 is merely an integrable and nonnegative function, we can approximate it from below in $L^1(\mathbf{R}^n)$ by compactly supported functions as in the previous result. The L^1-contraction property guarantees that the asymptotic limit also depends in a contractive way in L^1, hence the result.

THEOREM 2.5. *As $t \to \infty$ we have*

$$v(s,\phi,t) = \left[\frac{s}{(n-2)t}\right]^{-n/2}(1+o(1)) \tag{2.26}$$

uniformly in $\phi \in S^{n-1}$ and in s on sets of the form $\{s \geq (s_0+\varepsilon)\,t^{n/(n-2)}\}$, where $\varepsilon > 0$ may be arbitrarily small, while

$$v(s,\phi,t) = o(t^{-n/(n-2)}) \tag{2.27}$$

uniformly in $\phi \in S^{n-1}$ and in s on $\{s \leq (s_0-\varepsilon)\,t^{n/(n-2)}\}$.

3. Inner region. In this section we return to the radial solution $u(r,t)$ to derive an estimate in the inner region

$$0 \leq r \leq r_0(t) = \exp\{s_0 t^{n/(n-2)}\}, \quad s_0 = \kappa_0 M^{-2/(n-2)}. \tag{3.1}$$

Later on we extend the results to nonradial solutions.

We shall prove the following theorem in the radial case.

THEOREM 3.1. *Let $u(r,t)$ be the solution of the Cauchy problem* (1.1)–(1.3). *Then*

$$\log u(r,t) = -ns_0 t^{n/(n-2)}(1+o(1)) \quad \textit{as} \quad t\to\infty \tag{3.2}$$

uniformly in the inner region.

Proof. We proceed in several steps. Fix an $\varepsilon > 0$ small and denote

$$r_\varepsilon(t) = \exp\{(s_0+\varepsilon)\,t^{n/(n-2)}\}. \tag{3.3}$$

3.1. Interior regularity in the outer region. Set $I_\varepsilon = (s_0 + \varepsilon/2, s_0 + 3\varepsilon/2)$. We prove the following interior regularity result for (2.9).

PROPOSITION 3.2. *For $k = 1, 2, 3$ we have*

$$\left|\frac{\partial^k \theta}{\partial \xi^k}\right| \le C_k \quad \textit{for } \tau \gg 1 \textit{ and } \xi \in I_\varepsilon. \tag{3.4}$$

Proof. A gradient bound for $k = 1$ has already been proved; see (2.19)–(2.20). In order to prove a bound on the second derivative we note that in the domain $\xi \ge s_0+\varepsilon/4$ (where the limit function F_* is smooth) we can apply to (2.9) the classical Bernstein approach to prove the interior regularity. We will use the technique presented in [GV2, Proposition 5.4] for singular perturbations of first-order equations of the form (2.9). Let us review the main points and indicate the small differences. The right-hand side of (2.9) has two terms and the stationary operator $\mathbf{C}$ in the perturbation term is uniformly elliptic on the solution for $\tau \gg 1$ in $\xi \in I_\varepsilon$ and appears multiplied by an exponentially small factor. The first-order term preserves the regularity in the domain where $\theta_\xi < 0$ due to the entropy inequality. This is proved as in [GV2, p. 1125] by differentiating (2.9) with respect to ξ, performing a nonlinear change of the dependent variable $\theta_\xi = \varphi(v)$ (with a smooth function φ to be determined with the typical properties of the Bernstein method), differentiating again with respect to ξ, and setting

$$Z = \chi^2(\xi)(v_\xi)^2,$$

where the C^∞-function χ has to be chosen here with the following properties: it must be monotone and increasing, $\chi \equiv 0$ for all $\xi \le s_0 + \varepsilon/4$, and $\chi \equiv 1$ for $\xi \ge s_0 + \varepsilon/2$ (in other words, χ cuts off the shock; it does not cut off $\xi = +\infty$). We arrive at an equation of the form [GV2, (5.21)]

$$Z_\tau = A\, Z_\xi + B\, Z + e^{-\alpha\tau}\, J_1(Z)$$

with some coefficients A, B and a uniformly elliptic operator J_1. The important point is that J_1 is controlled in exactly the same way as for uniformly parabolic equations and the coefficient B satisfies $B \le 0$ in I_ε, $\tau \gg 1$. Since by the construction the function χ is not compactly supported from the right-hand side, we need some control at $\xi = +\infty$, but we may conclude that under the hypotheses on the initial data v_0, the second derivative $\theta_{\xi\xi}(\xi, \tau)$ is uniformly small for $\xi \gg 1$. Then a uniform in $\tau \gg 1$ estimate on Z in I_ε (which implies (3.4)) follows from the maximum principle as in [GV2, pp. 1126–1127].

The proof for $k = 3$ is similar. This completes the analysis. □

It is convenient now to consider the rescaled function

$$g = -\frac{n}{2}\,\theta^{-2/n} \tag{3.5}$$

("the pressure"). Then the rescaled equation (2.9) becomes

$$g_\tau = \alpha\Big(-\frac{2}{n}g + \xi g_\xi\Big) - \frac{2}{\alpha}g g_\xi + e^{-\alpha\tau}\mathbf{D}(g), \quad \mathbf{D}(g) = -\frac{2}{n}g g_{\xi\xi} + (g_\xi)^2. \tag{3.6}$$

As a consequence of Proposition 3.2, we have

$$\left|\frac{\partial^3 g}{\partial \xi^3}\right| \le C_3 \quad \text{for } \tau \gg 1 \text{ and } \xi \in I_\varepsilon. \tag{3.7}$$

3.2. Lateral analysis. From convergence (2.17) and regularity (3.7) we obtain an estimate near the lateral boundary of the inner domain of the pressure written in the original variables

$$w = -\frac{n}{2}\, u^{-2/n}. \tag{3.8}$$

It follows from (2.17) that in the outer region

$$w(r,t) = -\frac{n}{2(n-2)t}\, r^2 \log r\, (1+o(1)) \quad \text{as} \quad t \to \infty.$$

Since, in view of the interior regularity in the outer region, this asymptotic estimate can be differentiated in r twice, we arrive at the following result.

COROLLARY 3.3. *For $r = r_\varepsilon(t)$ we have*

$$\Delta w = -\psi_\varepsilon(t) \equiv -\frac{n^2}{n-2}(s_0+\varepsilon)\, t^{2/(n-2)}(1+o(1)) \quad \textit{as} \quad t \to \infty. \tag{3.9}$$

3.3. Inner semiconvexity. We now use the well-known semiconvexity approach [AB] in the inner region. Here and later on we denote by c, $c_1, \dots$ different positive constants.

PROPOSITION 3.4. *Let $\varepsilon > 0$. Then there exist a constant $c > 0$ and a $t_c > 0$ such that for $t > t_c$,*

$$\Delta w \geq -\tilde{\psi}_\varepsilon(t) \equiv -\psi_\varepsilon(t) - c \quad \textit{in the ball} \;\; B_\varepsilon(t) = \{r < r_\varepsilon(t)\}. \tag{3.10}$$

Proof. The pressure (3.8) solves in $\{w < 0\}$ the parabolic equation

$$w_t = -\frac{2}{n} w \Delta w + |\nabla w|^2.$$

Differentiating this equation twice we have that $z = \Delta w$ satisfies

$$z_t = -\frac{2}{n} w \Delta z + \frac{2(n-2)}{n} \nabla w \cdot \nabla z - \frac{2}{n} z^2 + 2 \sum_{(i,j)} \left(\frac{\partial^2 w}{\partial x_i \partial x_j} \right)^2.$$

As in [AB], using the Cauchy–Bunyakovskii–Schwarz inequality

$$\sum_{(i,j)} \left(\frac{\partial^2 w}{\partial x_i \partial x_j} \right)^2 \geq \frac{1}{n}\, z^2,$$

we arrive at a linear parabolic differential inequality of the form

$$z_t \geq -\frac{2}{n} w \Delta z + \frac{2(n-2)}{n} \nabla w \cdot \nabla z.$$

Therefore (3.10) follows from (3.9) by the maximum principle. □

3.4. Mass analysis: End of the proof. Integrating inequality (3.10) twice over $(0,r)$, we obtain the estimate

$$u(r,t) \geq u(0,t) \left(1 + \frac{r^2}{d^2(t)} \right)^{-n/2} \quad \text{in } B_\varepsilon(t), \tag{3.11}$$

where

$$d(t) = n\, u^{-1/n}(0,t)\tilde{\psi}_\varepsilon^{-1/2}(t).$$

Integrating (3.11) over $B_\varepsilon(t)$, we get the following estimate of the mass M_ε in the inner region:

(3.12)
$$M_\varepsilon(t) = n\omega_n \int_0^{r_\varepsilon(t)} r^{n-1}u(r,t)dr \ge n\omega_n u(0,t)d^n(t)\int_0^{l(t)} \eta^{n-1}(1+\eta^2)^{-n/2}d\eta,$$

with $l(t) = r_\varepsilon(t)/d(t)$. However,

$$\begin{aligned} M_\varepsilon(t) &= n\omega_n \int_{-\infty}^{(s_0+\varepsilon)t^{n/(n-2)}} v(s,t)\, ds \\ &= n\omega_n \int_{-\infty}^{s_0+\varepsilon} \theta(\xi,\tau)\, d\xi \\ &= n\omega_n(1+o(1))\int_{-\infty}^{s_0+\varepsilon} F_*(\xi)\, d\xi \quad \text{as} \quad t\to\infty. \end{aligned}$$

Therefore, the mass M_ε satisfies for small $\varepsilon > 0$

$$M_\varepsilon(t) = O(\varepsilon) \quad \text{as} \quad t\to\infty. \tag{3.13}$$

To conclude, we need to consider two cases. (i) Take a sequence $\{t_k\}\to\infty$ and assume that the sequence $\{l(t_k)\}$ is bounded. Then from (3.12) we obtain that for all $t = t_k \gg 1$,

$$M_\varepsilon(t) \ge c\, u(0,t)d^n(t)l^n(t) = c\, u(0,t)r_\varepsilon^n(t).$$

Therefore, it follows from (3.13) that $u(0,t)r_\varepsilon^n(t) \le c_1\,\varepsilon$ and hence

$$u(0,t) \le c_1\varepsilon r_\varepsilon^{-n}(t) = c_1\varepsilon \exp\{-n(s_0+\varepsilon)\, t^{n/(n-2)}\}.$$

Therefore we obtain an upper bound

$$\lim\sup_{t=t_k\to\infty} t^{-n/(n-2)}\log u(0,t) \le -n(s_0+\varepsilon). \tag{3.14}$$

(ii) Assume now that the sequence $\{l(t_k)\}$ is unbounded and without loss of generality $\{l(t_k)\}\to\infty$. Then one can calculate from (3.12) that for $t = t_k \gg 1$,

$$\begin{aligned} M_\varepsilon(t) &\ge n\omega_n\, u(0,t)d^n(t)\log l(t)(1+o(1)) \\ &= c\,\psi_\varepsilon^{-n/2}(t)\log l(t)(1+o(1)). \end{aligned} \tag{3.15}$$

As before, (3.13) yields that $\psi_\varepsilon^{-n/2}(t)\log l(t) \le c_1\varepsilon$, whence the estimate

$$\log l(t) = \frac{1}{n}\log u(0,t) + \log r_\varepsilon(t) + \frac{1}{2}\log\psi_\varepsilon(t) - c \le c_1\varepsilon\psi_\varepsilon^{n/2}(t) \le c_2\varepsilon\, t^{n/(n-2)}.$$

In view of (3.3) and (3.9) this implies that as $t = t_k\to\infty$ (cf. (3.14)),

$$\log u(0,t) \le -[ns_0 + O(\varepsilon)]t^{n/(n-2)}. \tag{3.16}$$

As a lower bound, it follows from the convergence given by (2.17) and from the eventual monotonicity property $u(r,t) \le u(0,t)$ for $t \gg 1$ (see the appendix) that

$$\log u(0,t) \ge \log u(r_\varepsilon(t),t) = -n(s_0+\varepsilon)t^{n/(n-2)}(1+o(1)). \tag{3.17}$$

From (3.16) (or (3.14)) and (3.17) we obtain uniform bounds from above and below of the solution in $B_\varepsilon(t)$ for $t \gg 1$. Since $\varepsilon > 0$ is arbitrarily small, we arrive at (3.2). □

3.5. The nonradial case. In order to extend the above inner radial analysis to the nonradial case, we just note that due to Theorem 2.5, we can bound (both above and below) the general solution $u(x,t)$ by radial ones:

$$\underline{u}(r,t) \le u(x,t) \le \overline{u}(r,t), \quad t \ge T,$$

where, by the outer analysis, the masses of $\underline{u}$ and $\overline{u}$ do not differ by more than 2ε. Indeed, the only thing we take from the outer analysis is the behavior of the solutions in a neighborhood of the inner lateral boundary $|x| = r_0(t)$ which is given by (2.26) for nonradial solutions.

4. Conclusions and final remarks.

1. We have rigorously established the two-region structure for the asymptotic behavior of finite-mass solutions of the fast-diffusion equation (1.1) with critical exponent $\sigma = 2/n$. This structure compares with the simpler one-region structure of noncritical exponents. The outer expansion has a *hyperbolic character* with natural variable $v = r^n u$, which develops an N-wave profile with a *shock* located at an exponentially growing distance $R(t)$, while it converges in relative size to the trivial state for $|x| < R(t)$. The main mathematical novelty in the proof of the hyperbolic structure is the use of entropy inequalities.

In the original variable and logarithmic scale we observe the formation of a *mesa-like profile.* Further analysis of the inner region should allow us to resolve the flatness of $\log u$ by calculating the second-order corrections which affect u as factors. The formal calculations are again given in [Ki] but a proof of those facts needs new techniques which fall outside the scope of this work.

2. We have studied the critical case in dimensions $n \ge 3$. Let us recall the situation in $n = 1, 2$, which is quite different. In dimension 1, where $\sigma = 2$, no finite-mass solutions exist, which eliminates the problem. For $n = 2$ we get the equation $u_t = \Delta \log u$. It is proved in [VER] that the Cauchy problem admits infinitely many solutions with finite mass and all of them extinguish in finite time, $T \le \|u_0\|_1/4\pi$, a striking contrast with the critical exponent for $n \ge 3$ discussed in this paper.

3. Explicit solutions are an important auxiliary tool in the investigation of nonlinear or asymptotic phenomena. Indeed, for $\sigma \ne 2/n$ the asymptotic behavior of finite-mass solutions has been described in terms of explicit self-similar solutions. No such solutions exist in the critical case, because of the nonexistence result of [BF]. However, we can exhibit self-similar solutions which represent the behavior of the outer zone. An explicit example is given by the function

$$u^{-2/n}(x,t) = \frac{|x|^2}{(n-2)t}\,\log\left(\frac{|x|}{at^\gamma}\right), \quad \gamma = \frac{n}{2(n-2)}, \tag{4.1}$$

where $a > 0$ is an arbitrary constant. The solution is defined only for $|x| > l(t) = a\,t^\gamma$, and blows up at $x = l(t)$. It has at infinity the correct behavior predicted by our results, since the rescaled function θ defined in (2.8) is given by

$$\theta^{-2/n}(\xi,\tau)=\frac{\xi}{n-2}-e^{-\alpha\tau}\left(\frac{s_0+\alpha\tau/2}{n-2}\right),\quad \alpha=2\gamma,\ \ s_0=\log a, \tag{4.2}$$

which is to be compared with F given by (2.14).

Appendix. We now prove some eventual monotonicity-like properties of radial solutions $u(r,t)$ and $v(r,t)$.

Proof of eventual monotonicity of $\boldsymbol{u(r,t)}$. According to the L^1-approximation scheme from section 2.5, it suffices to prove the property of eventual monotonicity for radial solutions $u(r,t)$ with sufficiently smooth compactly supported initial data

$$u_0(r)>0 \quad \text{for } r<R=2a/\varepsilon,\ \ u_0(r)=0 \text{ for } \ r\geq R. \tag{A.1}$$

Moreover, due to the approximation in L^1, we are free to impose a uniform slope condition on the approximating initial data $u_0\in C^1([0,R])$. Namely, we assume that

$$u_0'(R)=\mu>0, \tag{A.2}$$

where μ is chosen independent of the small approximating parameter ε.

We use the intersection comparison approach; see [GV4, p. 1100]. We compare $u(r,t)$ with a small flat solution $\tilde{u}\equiv\delta>0=\text{const}$ satisfying (1.1). By (A.1) and (A.2), $u_0(r)$ intersects the level δ exactly once. By the Sturmian argument, the number of intersections $J(t)$ does not increase so that

$$J(t)\leq 1 \quad \text{for all } t>0. \tag{A.3}$$

Since

$$u(r,t)\to 0 \quad \text{as } t\to\infty \text{ uniformly}, \tag{A.4}$$

there exists a moment $t_1>0$ such that $J(t)=1$ for all $t\in[0,t_1)$ and $J(t)=0$ for $t\geq t_1$. By the strong maximum principle (this admits a simple geometric interpretation; see [GV4]), for $t>t_1$ the solution $u(r,t)$ of a uniformly parabolic equation is strictly monotone in r.

Proof of property $\boldsymbol{(M)}$ for $\boldsymbol{v(r,t)}$. We again restrict ourselves to radial compactly supported data satisfying (A.1) and (A.2). We now apply the intersection-comparison idea from [GV4] to the radial equation (2.2).

We compare $v(r,t)$ with the stationary solutions $\tilde{v}=\delta>0$. In terms of the dependent variable u this corresponds to the comparison with the singular stationary solution $\tilde{u}=\delta r^{-n}$. It follows from (A.2) that there exists a small $\delta>0$ such that the constant solution $\tilde{v}=\delta$ intersects $v_0(r)$ exactly once for $r\approx R-0$. By continuity, the same is true for any small $t=\tau>0$. On the other hand, in terms of the corresponding solutions $u(r,t)$ and $\tilde{u}=\delta r^{-n}$ the profiles $u(r,\tau)$ ($\tau>0$ is a small time-shift which establishes a natural interior regularity of the solution) and $\tilde{u}(r)$ intersect each other exactly once for $r\approx 0$ provided that $\delta\ll 1$. Finally, we conclude that there exists $\delta>0$ sufficiently small such that the number of intersections $J(\tau)$ between the profiles $u(r,\tau)$ and $\tilde{u}(r)$ satisfies $J(\tau)=2$. Hence, by the Sturmian argument,

$$J(t)\leq 2,\quad t\geq\tau. \tag{A.5}$$

Since $v(r,t)$ satisfies the asymptotic property (A.4), the estimate above (A.5) means that there exists a moment t_1 such that $J(t_1) = 0$, $J(t) = 2$ for $t < t_1$ (and $J(t) = 0$ for all $t > t_1$ by the standard comparison). As in [GV4, section 10], using a simple geometric interpretation based on the application of the strong maximum principle for uniformly parabolic equations, one concludes that all the nonmonotonicities of $v(r,t)$ must be destroyed at this moment $t = t_1$, and the profile $v(r,t)$ has exactly one maximum and no minima for all $t > t_1$. This mimics the assumption (M) which happens eventually in time.

Acknowledgments. We would like to thank J. R. King for useful discussions. This paper was started while the first author was a visiting professor at the Department of Mathematics, University Autónoma of Madrid, and then as “Profesor Visitante Iberdrola de Ciencia y Tecnología.” He is grateful to both institutions for their hospitality and support.

REFERENCES

[AB] D. G. Aronson and P. Bénilan, *Régularité des solutions de l'équation des milieux poreux dans* $\mathbf{R^n}$, C. R. Acad. Sci. Paris. Ser. I, 288 (1979), pp. 103–105.

[B] G. I. Barenblatt, *On some unsteady motions of a liquid and a gas in a porous medium*, Prikl. Mat. Mekh., 16(1) (1952), pp. 67–78.

[BC] Ph. Bénilan and M. G. Crandall, *The continuous dependence on* φ *of solutions of* $u_t - \Delta\varphi(u) = 0$, Indiana Univ. Math. J., 30 (1981), pp. 161–177.

[BF] H. Brezis and A. Friedman, *Nonlinear parabolic equations involving measures as initial conditions*, J. Math. Pures Appl., 62 (1983), pp. 73–77.

[CVW] L. A. Caffarelli, J. L. Vazquez, and N. I. Wolanski, *Lipschitz-continuity of solutions and interfaces of the N-dimensional porous medium equation*, Indiana Univ. Math. J., 36 (1987), pp. 373–401.

[EVZ1] M. Escobedo, J. L. Vazquez, and E. Zuazua, *Asymptotic behavior and source-type solutions for a diffusion-convection equation*, Arch. Rational Mech. Anal., 124 (1993), pp. 43–65.

[EVZ2] M. Escobedo, J. L. Vazquez, and E. Zuazua, *A diffusion-convection equation in several space variables*, Indiana Univ. Math. J., 42 (1993), pp. 1413–1440.

[FK] A. Friedman and S. Kamin, *The asymptotic behavior of gas in an n-dimensional porous medium*, Trans. Amer. Math. Soc., 262 (1980), pp. 551–563.

[GP] V. A. Galaktionov and L. A. Peletier, *Asymptotic behaviour near finite time extinction for the fast diffusion equation*, Arch. Rational Mech. Anal., 139 (1997), pp. 83–98.

[GV1] V. A. Galaktionov and J. L. Vazquez, *Asymptotic behaviour of nonlinear parabolic equations with critical exponents. A dynamical systems approach*, J. Funct. Anal., 100 (1991), pp. 435–462.

[GV2] V. A. Galaktionov and J. L. Vazquez, *Extinction for a quasilinear heat equation with absorption* II. *A dynamical systems approach*, Comm. Partial Differerential Equations, 19 (1994), pp. 1107–1137.

[GV3] V. A. Galaktionov and J. L. Vazquez, *A dynamical systems approach for the asymptotic analysis of nonlinear heat equations*, in International Conference on Differential Equations, Proceedings Equadiff-95 Conference, Lisboa, 1996, L. Magalhaes et al., eds., World Scientific, Singapore, 1998, pp. 82–106.

[GV4] V. A. Galaktionov and J. L. Vazquez, *Extinction for a quasilinear heat equation with absorption* I. *Technique of intersection comparison*, Comm. Partial Differerential Equations, 19 (1994), pp. 1075–1106.

[GNN] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[Ki] J. R. King, *Self-similar behaviour for the equation of fast nonlinear diffusion*, Philos. Trans. Roy. Soc. London Ser. A, 343 (1993), pp. 337–375.

[Kr] S. N. Kruzhkov, *First-order quasilinear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.

[LL] L. D. Landau and E. M. Lifschitz, *Fluid Mechanics*, Pergamon Press, Oxford, 1959.

[LP] T.-P. Liu and M. Pierre, *Source-solutions and asymptotic behavior in conservation laws*, J. Differential Equations, 51 (1984), pp. 419–441.

[PZ] M. A. Peletier and H. Zhang, *Self-similar solutions of a fast diffusion equation that do not conserve mass*, Differential Integral Equations, 8 (1995), pp. 2045–2064.

[S] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, Berlin, 1983.

[V] J. L. Vazquez, *Nonexistence of solutions for nonlinear heat equations of fast diffusion type*, J. Math. Pures Appl., 71 (1992), pp. 503–526.

[VER] J. L. Vazquez, J. R. Esteban, and A. Rodriguez, *The fast diffusion equation with logarithmic nonlinearity and the evolution of conformal metrics in the plane*, Adv. Differential Equations, 1 (1996), pp. 21–50.

[ZK] Ya. B. Zel'dovich and A. S. Kompaneetz, *Towards a theory of heat conduction with thermal conductivitydepending on the temperature*, in Collection of Papers Dedicated to 70th Birthday of Academician A. F. Ioffe, Izd. Akad. Nauk SSSR, Moscow, 1950, pp. 61–72.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1175–1191

INTERFACE BEHAVIOR OF COMPRESSIBLE NAVIER–STOKES EQUATIONS WITH VACUUM*

TAO LUO[†], ZHOUPING XIN[‡], AND TONG YANG[§]

Abstract. In this paper, we study a one-dimensional motion of viscous gas near vacuum with (or without) gravity. We are interested in the case that the gas is in contact with the vacuum at a finite interval. This is a free boundary problem for the one-dimensional isentropic Navier–Stokes equations, and the free boundaries are the interfaces separating the gas from vacuum, across which the density changes continuously. The regularity and behavior of the solutions near the interfaces and expanding rate of the interfaces are studied. Smoothness of the solutions is discussed. The uniqueness of the weak solutions to the free boundary problem is also proved.

Key words. interface, Navier–Stokes equations, vacuum

AMS subject classification. 35Q10

PII. S0036141097331044

1. Introduction.

We consider the evolution of the interfaces separating one-dimensional isentropic viscous gases from vacuum when the gases are in contact with the vacuum on a finite interval initially, with (or without) external force. The important feature of this problem is that the density changes continuously across the interfaces separating the gases and vacuum. This models many interesting phenomena, such as gaseous stars problems in astrophysics [9]. For further physical significance and mathematical treatment of such free boundaries, we refer to the excellent survey paper of Nishida [11].

The one-dimensional isentropic viscous gas flow is governed by the compressible Navier–Stokes equations which can be rewritten, in Eulerian coordinates, as (in the case without external force)

$$\begin{aligned} &\rho_t + (\rho u)_x = 0, \\ &(\rho u)_t + (\rho u^2 + P(\rho))_x = \mu u_{xx}, \end{aligned} \tag{1.1}$$

where $x \in \mathbf{R^1}$ and $t > 0$, and $\rho = \rho(x,t)$, $u = u(x,t)$, and $P(\rho)$ denote, respectively, the density, velocity, and the pressure; $\mu > 0$ is the viscosity constant. For simplicity of presentation, we consider only the polytropic gas, i.e., $P(\rho) = A\rho^\gamma$ with $\gamma > 1$, $A > 0$ being constants.

We consider (1.1) with the initial data

$$\rho(x,0) = \rho_0(x), \qquad u(x,0) = u_0(x). \tag{1.2}$$

*Received by the editors December 4, 1997; accepted for publication (in revised form) February 25, 1999; published electronically May 11, 2000.

http://www.siam.org/journals/sima/31-6/33104.html

[†] Department of Mathematics, City University of Hong Kong. Current address: Institute of Mathematics, Academia Sinica, China. The research of this author was partially supported by a small scale grant of City University of Hong Kong 9030642 and RGC Competitive Earmarked research grant 9040190.

[‡] Courant Institute, New York University. Current address: IMS and Department of Mathematics, The Chinese University of Hong Kong. The research of this author was partially supported by NSF grant DMS-96-00137 and DOE grant DE-FG02-88ER-25053.

[§] Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong (matyang@cityu.edu.hk). The research of this author was partially supported by the RGC Competitive Earmarked research grant 9040190.

Our main assumption is that the entire gas initially occupies a finite interval $(a, b) \subset \mathbf{R^1}$ and is in contact with the vacuum. More precisely, the initial density $\rho_0(x)$ is supposed to satisfy

(A1) $\rho_0(x) \in C(\mathbf{R^1})$, supp$\rho_0 \subset (a, b)$, $\rho_0(x) > 0\ \forall x \in (a, b)$, and

(A2) $\rho_0^k \in H^1([a, b])$ for some constant $0 < k \leq \gamma - 1/2$.

The initial velocity, $u_0(x)$, is assumed to possess the following regularity:

(A3) $\rho_0 u_0^2$, $(\partial_x u_0)^2$, $\rho_0^{-1}(\partial_{xx} u_0)^2 \in L^1([a, b])$.

It is expected that velocity will be smooth enough up to the interfaces which separate the gas from the vacuum so that the interfaces are particle paths. Suppose $x = a(t)$ and $b = b(t)$ are two particle paths issuing from a and b, respectively, i.e.,

$$\dot{a}(t) = u(a(t), t),\ \dot{b}(t) = u(b(t), t) \tag{1.3}$$

with $a(0) = a$ and $b(0) = b$. The free boundary problem to be studied is

$$\begin{aligned} &(1.1),\ \text{in } (a(t), b(t)) \times (0, +\infty), 0 < t < +\infty,\\ &\rho(a(t), t) = \rho(b(t), t) = 0,\\ &(\rho, u)(x, 0) = (\rho_0, u_0)(x), x \in [a, b]. \end{aligned} \tag{1.4}$$

The definition of weak solutions for this free boundary problem is given by the following definition.

Definition 1.1. *A function* (ρ, u) *is called a weak solution to the free boundary problem* (1.4) *if there exist* $a(t), b(t) \in C[0, \infty)$ *such that* $\rho \in C[(0, \infty), L^2([a(t), b(t)])]$, $u \in C[(0, \infty), H^1([a(t), b(t)])]$, *and* $\lim_{x \to a(t)+} \rho(x, t) = 0 = \lim_{x \to b(t)-} \rho(x, t)$. *Furthermore,*

$$\begin{aligned} &\int_a^b \rho_0 v(x, 0) dx + \int_0^{+\infty} \int_{a(t)}^{b(t)} (\rho v_t + \rho u v_x) dx dt = 0,\\ &\int_a^b \rho_0 u_0 w(x, 0) dx + \int_0^{+\infty} \int_{a(t)}^{b(t)} (\rho u w_t + (\rho u^2 + P(\rho) + \mu u_x) w_x) dx dt = 0 \end{aligned}$$

hold for all test functions v, $w \in C_0^1(\Omega)$ *with* $\Omega = \{(x, t) | a(t) \leq x \leq b(t), 0 \leq t < +\infty\}$.

In what follows, we always use C $(C(T))$ to denote a generic positive constant depending only on the initial data (and the given time T). We now state our first result on the existence and behavior of weak solutions to the free boundary problem (1.4).

Theorem 1.2. *Let* ρ_0 *and* u_0 *satisfy* (A1)–(A3). *Then the free boundary problem* (1.4) *admits a globally defined weak solution with* $a(\cdot), b(\cdot) \in C^1[0, +\infty)$. *Moreover, the solution* $(\rho, u)(x, t)$ *has the following properties:*

(1) *Regularity of the solution.*

$\rho > 0$ *in* Q *with* $Q =: \{(x,\ t) : a(t) < x < b(t), 0 \leq t < +\infty\}$.

$$\int_{a(t)}^{b(t)} (\rho u^2 + u_x^2)(x, t) dx \leq C, \qquad 0 \leq t < +\infty, \tag{1.5}$$

$$\int_{a(t)}^{b(t)} ([(\rho^k)_x]^2 + \rho^{-1} u_{xx}^2)(x, t) dx \leq C(T), \qquad 0 \leq t \leq T, \tag{1.6}$$

for any $T > 0$.

(2) *Decay rate of the density and expanding rate of the interface.* *There exist positive constants c_i $(i = 1, 2, \ldots, 7)$, independent of the time t, such that*

$$c_1\rho_0(x_1)(1 + c_2\rho_0^\gamma(x_1)t)^{-1/\gamma} \le \rho(x,t) \le c_3\rho_0(x_1)(1 + c_4\rho_0^\gamma(x_1)t)^{-1/\gamma}, \tag{1.7}$$

where x_1 is determined uniquely by $\int_a^{x_1}\rho_0(z)dz = \int_{a(t)}^x \rho(z,t)dz$ for any x,

$$c_5(1+t)^{1/\gamma} \le b(t) - a(t) \le c_6(1+t)^{1/\gamma}, \qquad 0 \le t < \infty, \tag{1.8}$$

$$\sup\nolimits_{a(t)\le x\le b(t)}|u(x,t)| \le c_7(1+t)^{1/\gamma}, \qquad 0 \le t < \infty. \tag{1.9}$$

(3) *Behavior near the interface.*

$$\rho(x,t) \le C(T)|x - a(t)|^{1/2k}, \;\; \rho(x,t) \le C(T)|x - b(t)|^{1/2k}, \tag{1.10}$$

$$|u_x(x,t) - u_x(a(t)+,t)| \le C(T)|x - a(t)|^{(\frac{1}{4k}+\frac{1}{2})}, \tag{1.11}$$

$$|u_x(x,t) - u_x(b(t)-,t)| \le C(T)|x - b(t)|^{(\frac{1}{4k}+\frac{1}{2})} \tag{1.12}$$

for $0 \le t \le T$.

(4)

$$\int_{a(t)}^x (\rho u_t + \rho u u_x)(z,t)dx + P(\rho) = \mu u_x$$

for $a(t) < x < b(t)$ and $t > 0$, and

$$\int_{a(t)}^{b(t)} \rho u(x,t)dx = \int_a^b \rho_0 u_0(x)dx \tag{1.13}$$

for $t > 0$.

Remark 1.3. Part (3) of Theorem 1.2 is a simple consequence of (1.6). Indeed,

$$\rho^k(x,t) = \int_{a(t)}^x (\rho^k)_x dx \le \left(\int_{a(t)}^x [(\rho^k)_x]^2 dx)^{1/2}|x - a(t)|^{1/2} \le C(T)|x - a(t)|^{1/2}\right),$$

which implies (1.10). Moreover, the Hölder inequality and (1.10) give

$$\begin{aligned}
&|u_x(x,t) - u_x(a(t)+,t)| = \left|\int_{a(t)}^x u_{xx}\right| \\
&\le \left(\int_{a(t)}^x \rho^{-1}u_{xx}^2\right)^{1/2}\left(\int_{a(t)}^x \rho\right)^{1/2} \\
&\le C(T)\left(\int_{a(t)}^x (x - a(t))^{1/2k}\right)^{1/2} \\
&\le C(T)(x - a(t))^{(\frac{1}{4k}+\frac{1}{2})}.
\end{aligned}$$

Remark 1.4. It can be verified that the conservation of momentum property (1.13) holds true for the general weak solution defined in Definition 1.1 under the assumption (A_1)–(A_3) (see [10]).

Our next theorem shows that the solutions obtained in Theorem 1.2 are indeed smooth in the region $\{(x,t) : \rho(x,t) > 0\}$ if the initial data have the appropriate regularity, and the smoothness is up to the boundary if the initial density is connected to vacuum very smoothly. Precisely, we have the following theorem.

THEOREM 1.5. *Let* (ρ, u) *be the weak solution to* (1.4) *described in Theorem* 1.2. *Then it holds that*

(1) *if* $\rho_0^\alpha(\partial_x^2\rho_0)^2 \in L^1[a,b]$, *then* $\rho^{\alpha_1}(\partial_x^2\rho)^2(\cdot,t) \in L^1[a(t),b(t)]$ *with* $\alpha_1 = 2k-3+\max\{2k+3+\alpha, 6k+1\}$. *And the weak solution* (ρ,u) *to* (1.4) *is smooth in the region* $Q =: \{(x,\ t) : a(t) < x < b(t), 0 \le t < +\infty\}$ *with* $\rho(\cdot,t) \in C^{1+\lambda}(a(t),b(t))$, $u(\cdot,t) \in C^{2+\lambda}(a(t),b(t))$ *for some* $0 < \lambda < 1$ *and any* $t > 0$;

(2)

$$u_{xx}(a(t)+,t) = u_{xx}(b(t)-,t) = 0$$

if $k < \frac{1}{2}$; *furthermore*

$$\rho(\cdot,t) \in H^2[a(t),b(t)], \qquad 0 \le t \le \infty, \tag{1.14}$$

if $\alpha \le -4k$ *and* $k \le 1/4$, *and this implies* ρ_x *is Hölder continuous in the whole interval* $[a(t),b(t)]$ *for any* $t \ge 0$. *Also, the gradient of pressure satisfies*

$$|P(\rho)_x(x,t)| \le C(T)\rho^{\gamma-1} \le C(T)(|x-a(t)|^{(\gamma-1)/2k} + |x-b(t)|^{(\gamma-1)/2k}). \tag{1.15}$$

Finally, we have the following uniqueness result.

THEOREM 1.6 (uniqueness). *Assume* (A_1)–(A_3), *let* (ρ_1,u_1) *and* (ρ_2,u_2) *be two weak solutions to the free boundary problem* (1.4) *in* $0 \le t \le T$ *as described in Definition* 1.1. *Then* $(\rho_1,u_1)(x,t) = (\rho_2,u_2)(x,t)$ *in* $a(t) < x < b(t)$ *and* $0 \le t \le T$.

Remark 1.7. This uniqueness result particularly implies that the whole sequence of approximate solutions constructed in section 2 converges to a unique weak solution.

Remark 1.8. So far, all of our results are stated for the case without external force. However, one can check by modifying our analysis slightly that similar results hold true in the presence of external forces such as gravity.

The free boundary problem of one-dimensional Navier–Stokes equations with one boundary fixed was investigated by Okada in [12]; see also [11], where the global existence of the weak solutions was proved and the regularity of $(\rho^{1/2}u, u_x, \rho^{-1/2}u_{xx})(\cdot,t) \in L^2, \rho \in BV$ was obtained. Similar results were derived in [10], [13] for the equations of spherically symmetric motion of viscous gases. It also should be noticed that another interesting class of free boundary problem of viscous gases for the one-dimensional viscous gases which expands into the vacuum has been studied by many people; see [1], [2], [3], [8].

It is important and interesting to investigate the regularity of the solution to the above free boundary problem and the behavior near interfaces due to the degeneracy of vacuum. In general, we do not expect to prove the global existence of the smooth solution to the above free boundary problem because of the degeneracy. However, as we showed in Theorem 1.5, if the initial density is connected to vacuum smoothly enough, the smoothness of the solution can be up to the boundary. Also we get the clear description of the behavior of solutions near the interfaces between the gas and vacuum. For these, the first observation is that $(\rho^k)_x(\cdot,t) \in L^2$, which indicates how the density is connected to the vacuum according to the value of k. The second ingredient in our analysis to improve the regularity of the solution is an L^2-estimate of $\rho^{\alpha_1/2}\rho_{xx}$.

It should be noted that Xin [16] proved recently that the smooth solution $(\rho, u) \in H^3(\mathbf{R^1})$ of the Cauchy problem of Navier–Stokes equations with compact density must blow up in the finite time. However, the different phenomena occur for the free boundary problem (1.4). As we showed in Theorem 1.2, the smooth solution of the free boundary problem can survive for all time if the initial density is connected to vacuum very smoothly. In fact, $(\rho, u) \in H^3(\mathbf{R^1})$ implies $u = 0$ outside the interfaces [16] for the Cauchy problem. Hence the interfaces separating the gas and vacuum must be fixed, i.e., independent of the time. But the estimate (1.8) shows the interfaces for the free boundary problem (1.4) are not fixed.

We should mention that the free boundary problem of a modified Navier–Stokes equation was studied by Liu, Xin, and Yang in [8] with the boundary condition $u_x = P(\rho)$, where the initial density was assumed to connect vacuum with discontinuities, i.e., $inf_{x\in(a,b)}\rho_0(x) \geq \delta > 0$, $\rho_0(x) = 0$ when $x < a$ or $x > b$. This property can be maintained for some finite time [8], and the local existence of the weak solution was proved then.

The important progress has been made on compressible Navier–Stokes equations when the initial density is away from vacuum in several aspects, for smooth initial data or discontinuous initial data, one-dimensional or multidimensional problem. For these results, please refer to [1], [2], [3], [4], [5], [6], [14], [15], and [17].

The rest of this paper is organized as follows. In section 2, we convert the free boundary problem to a fixed boundary problem by using Lagrangian coordinates and giving some basic estimates. Based on these estimates, we study the smoothness of solutions and complete the proof of Theorem 1.5 with the L^2-estimate of $\rho^{\alpha_1/2}\rho_{xx}$ in section 3. The uniqueness result will be proved in section 4.

2. Global existence of weak solutions and the basic estimates. To solve the free boundary problem (1.4), it is convenient to convert the free boundaries to the fixed boundaries by using Lagrangian coordinates. Using the following coordinates transformation

$$y = \int_{a(t)}^{x} \rho(z,t)dz, \ \tau = t,$$

the free boundaries $x = a(t)$ and $x = b(t)$ become $\hat{a}(\tau) = 0$ and $\hat{b}(\tau) = \int_{a(t)}^{b(t)} \rho(x,t)dx = \int_a^b \rho_0(x)dx$, where $\int_a^b \rho_0(x)dx$ is the total mass initially, and without loss of generality, we normalize it to be 1. So in terms of Lagrangian coordinates, the free boundary problem (1.4) becomes

$$\begin{aligned} &\rho_\tau + \rho^2 u_y = 0, \\ &u_\tau + P(\rho)_y = (\mu\rho u_y)_y, 0 < y < 1, \tau > 0, \end{aligned} \tag{2.1}$$

with the data given by

$$\begin{aligned} (\rho, u)(y, 0) = (\rho_0(y), u_0(y)), 0 \leq y \leq 1, \\ \rho(0,t) = \rho(1,t) = 0, t > 0. \end{aligned} \tag{2.2}$$

The assumptions corresponding to (A1)–(A3) are transformed into

(A1′) $\rho_0(y) > 0$ as $y \in (0,1)$;

(A2′) $\rho_0^q \in H^1([0,1])$ with $1/2 < q = k + 1/2 \le \gamma$;
and

(A3′) $u_0, \rho_0^{1/2} u_{0y}, (\rho_0 u_{0y})_y \in L^2[0,1]$.
In addition,

$$0 < \int_0^1 (\rho_0)^{-1}(y)dy = \int_a^b 1dx = b - a < \infty. \tag{2.3}$$

It is more convenient to use t instead of τ in the case without confusion. Similarly we have the definition of weak solutions in Lagrangian coordinates corresponding to Definition 1.1.

To prove Theorem 1.2, we first prove the following Theorem 2.1.

Theorem 2.1. *Suppose* (A1′)–(A3′) *and* (2.3) *are satisfied. Then the initial-boundary value problem* (2.1) *and* (2.2) *admits a globally defined weak solution* (ρ, u) (y,t) *in* $[0,1] \times (0,\infty)$ *with the following properties:*

(1)

$$c_1\rho_0(y)(1 + c_2\rho_0^\gamma(y)t)^{-1/\gamma} \le \rho(y,t) \le c_3\rho_0(y)(1 + c_4\rho_0^\gamma(y)t)^{-1/\gamma}, 0 \le y \le 1, t > 0 \tag{2.4}$$

for some positive constant c_i $(i = 1, \ldots, 4)$ *independent of* t;

(2)

$$\int_0^1 (u^2 + \rho u_y^2 + u_t^2)(y,t)dy + \int_0^\infty \int_0^1 (\rho u_y^2 + u_t^2 + \rho u_{yet}^2)dydt \le C, 0 \le t < +\infty, \tag{2.5}$$

$$\max_{y\in[0,1]}|\rho u_y(y,t)| + \int_0^1 ([(\rho^q)_y]^2 + [(\rho u_y)_y]^2) + |u_y|)(y,t)dy \le C(T), 0 \le t \le T, \tag{2.6}$$

$$c_5(1+t)^{1/\gamma} \le \int_0^1 \rho^{-1}(y,t)dy \le c_6(1+t)^{1/\gamma},\ \max_{y\in[0,1]}|u(y,t)| \le C(1+t)^{1/\gamma}, t > 0, \tag{2.7}$$

for some positive constants c_5 *and* c_6 *independent of* t *and* $T > 0$ *is given;*

(3)

$$\int_0^y u_t(z,t)dz + P(\rho) = \rho u_y,$$

$$\int_0^1 u(y,t)dy = \int_0^1 u_0(y)dy.$$

To prove Theorem 2.1, we will construct the global solution to the initial boundary problem (2.1) and (2.2) by a slight modification of the method of lines used in [1], [8], [11], [12], [13], which can be described as follows. For simplicity, we take $\mu = 1$. For any given positive integer N, let $h = 1/N$. Consider the system of $2N$ ordinary differential equations

$$\begin{aligned} &\frac{d}{dt}\rho_{2n}^h + (\rho_{2n}^h)^2 \frac{u_{2n+1}^h - u_{2n-1}^h}{h} = 0,\\ &\frac{d}{dt}u_{2n-1}^h + \frac{P(\rho_{2n}^h) - P(\rho_{2n-2}^h)}{h}\\ &= \frac{1}{h^2}\{\rho_{2n}^h(u_{2n+1}^h - u_{2n-1}^h) - \rho_{2n-2}^h(u_{2n-1}^h - u_{2n-3}^h)\}, \end{aligned} \tag{2.8}$$

where $n = 1, 2, 3 \cdots N$, with the boundary conditions

$$\rho_0^h = \rho_{2N}^h = 0, \tag{2.9}$$

and the initial conditions

$$\rho_{2n}^h(0) = \rho_0\left(2n \cdot \frac{h}{2}\right), u_{2n-1}^h(0) = u_0\left((2n-1)\cdot\frac{h}{2}\right). \tag{2.10}$$

For $n = 1$ and N, we set $u_{-1}^h = u_{2N+1}^h = 0$.

In the following, we will use (ρ_{2n}, u_{2n-1}) instead of $(\rho_{2n}^h, u_{2n-1}^h)$ when it does not cause any confusion. For simplicity of presentation, we use $\sum$ instead of $\sum_{n=1}^N$ unless otherwise stated.

We now begin to derive some bounds on the solutions to (2.8)–(2.10). At first, standard energy estimate gives the following lemma.

Lemma 2.2. *Let $(\rho_{2n}(t), u_{2n-1}(t))$, $n = 1, 2, \ldots, N$ be the solution for* (2.8), (2.9), *and* (2.10). *Then we have*

$$\begin{aligned} &\sum\left(\frac{1}{2}u_{2n-1}^2(t) + \int_0^{\rho_{2n}(t)} s^{-2}P(s)ds\right)h\\ &\quad + \int_0^t \sum \rho_{2n}\left(\frac{u_{2n+1} - u_{2n-1}}{h}\right)^2 hds\\ &= \sum\left(\frac{1}{2}u_{2n-1}^2(0) + \int_0^{\rho_{2n}(0)} s^{-2}P(s)ds\right)h. \end{aligned} \tag{2.11}$$

Consequently, problem (2.8)–(2.10) has a unique global solution. In the next lemma, we give the decay rate of density function. The main idea for this is to get ordinary differential equations governing the density function along the particle path.

Lemma 2.3. *There exist positive constants c_i $(i = 1, \ldots, 4)$ independent of t and n such that*

$$c_1\rho_{2n}(0)(1 + c_2\rho_{2n}^\gamma(0)t)^{-1/\gamma} \le \rho_{2n}(t) \le c_3\rho_{2n}(0)(1 + c_4\rho_{2n}^\gamma(0)t)^{-1/\gamma}, n = 1, 2\cdots N, \tag{2.12}$$

for $0 \le t \le \infty$.

Proof. It follows from $(2.8)_1$ that

$$\rho_{2n}(t) = \rho_{2n}(0)\exp\left(-\int_0^t \frac{\rho_{2n}(u_{2n+1}-u_{2n-1})}{h}ds\right)$$

for $n = 1, 2, \ldots, N$, $t \geq 0$.

On the other hand, in view of $(2.8)_2$ and the boundary conditions (2.9), one has

$$\begin{aligned}
&\frac{\rho_{2n}(u_{2n+1}-u_{2n-1})}{h}\\
&=\sum_{j=1}^{n}\left[\rho_{2j}\frac{(u_{2j+1}-u_{2j-1})}{h}-\rho_{2j-2}\frac{(u_{2j-1}-u_{2j-3})}{h}\right]\\
&=\sum_{j=1}^{n}\left[\frac{d}{dt}u_{2j-1}h+(P(\rho_{2j})-P(\rho_{2j-2}))\right]\\
&=\sum_{j=1}^{n}\left(\frac{d}{dt}u_{2j-1}h\right)+P(\rho_{2n}). \qquad (2.13)
\end{aligned}$$

Thus,

$$\rho_{2n}(t) = \rho_{2n}(0)\exp\left(\sum_{j=1}^{n}(u_{2j-1}(0)-u_{2j-1}(t))h\right)\exp\left(-\int_0^t P(\rho_{2n}(s))ds\right). \qquad (2.14)$$

Since $P(\rho) = A\rho^\gamma$, one gets from (2.14) that

$$\begin{aligned}
&\frac{d}{dt}\left(\frac{1}{\gamma}\exp\left(\gamma\int_0^t P(\rho_{2n}(s))ds\right)\right)\\
&=A\rho_{2n}^{\gamma}(0)\exp\left(\gamma\sum_{j-1}^{n}(u_{2j-1}(0)-u_{2j-1}(t))h\right).
\end{aligned}$$

Integrating this over $[0,t]$ yields

$$\begin{aligned}
&\exp\left(\gamma\int_0^t P(\rho_{2n}(s))ds\right)\\
&=1+A\gamma\rho_{2n}^{\gamma}(0)\int_0^t\exp\left(\gamma\sum_{j-1}^{n}(u_{2j-1}(0)-u_{2j-1}(s))h\right)ds.
\end{aligned}$$

This, together with (2.14), leads to

$$\begin{aligned}
&\rho_{2n}(t)\\
&=\rho_{2n}(0)\exp\left(\sum_{j=1}^{n}(u_{2j-1}(0)-u_{2j-1}(t))h\right)\\
&\cdot\left\{1+A\gamma\rho_{2n}^{\gamma}(0)\int_0^t\exp\left(\gamma\sum_{j=1}^{n}(u_{2j-1}(0)-u_{2j-1}(s))h\right)ds\right\}^{-1/\gamma}. \qquad (2.15)
\end{aligned}$$

Equation (2.12) immediately follows from (2.11) and (2.15) and the Cauchy inequality. □

The next two lemmas yield uniform estimates on the approximations to derivatives.

Lemma 2.4.

$$\sum_{n=1}^{N}\rho_{2n}(\frac{u_{2n+1}-u_{2n-1}}{h})^2(t)h+\int_0^t\sum_{n=1}^{N}\dot{u}_{2n-1}^2(s)hds\le C; \tag{2.16}$$

hereafter, $\dot{f}$ denotes $\frac{df}{dt}$ for any function f of t.

Proof. Set $A_n(t)=(\rho_{2n})^{1/2}(\frac{u_{2n+1}-u_{2n-1}}{h})(t)$ and $B_n(t)=(\rho_{2n})(\frac{u_{2n+1}-u_{2n-1}}{h})(t)$ for $0\le t<+\infty$, $n=1,2,\dots,N$. A direct calculation shows that

$$\begin{aligned}&\sum A_n^2(t)h+2\sum[P(\rho_{2n})(\rho_{2n}^{-1/2})](t)A_n(t)h+2\int_0^t\sum(\dot{u}_{2n-1}^2+P'(\rho_{2n})B_n^2)(s)hds\\&=\sum A_n^2(0)h+2\sum[P(\rho_{2n})(\rho_{2n}^{-1/2})](0)A_n(0)h-\int_0^t\sum B_n(s)A_n^2(s)hds.\end{aligned} \tag{2.17}$$

It is routine to get (cf. [12])

$$\sum A_n^2(t)h+\int_0^t\sum\dot{u}_{2n-1}^2hds\le C+\int_0^t\left(\sum A_n^2(s)h\right)^2ds. \tag{2.18}$$

Since $\int_0^t\sum A_n^2(s)hds\le C$ (cf. Lemma 2.2), (2.16) follows from Gronwall's inequality. □

Lemma 2.5.

$$\sum\left(\frac{\rho_{2n}^q-\rho_{2n-2}^q}{h}\right)^2(t)h\le C(T) \tag{2.19}$$

for $0\le t\le T$, where q is the constant in (A2′).

Proof. It follows from (2.8) that

$$\begin{aligned}&\left(\rho_{2n}^q\sum_{j=1}^{n}\dot{u}_{2j-1}-\rho_{2n-2}^q\sum_{j=1}^{n-1}\dot{u}_{2j-1}\right)+\frac{P(\rho_{2n})\rho_{2n}^q-P(\rho_{2n-2})\rho_{2n-2}^q}{h}\\&=-\frac{1}{q}\frac{d}{dt}\left(\frac{\rho_{2n}^q-\rho_{2n-2}^q}{h}\right).\end{aligned} \tag{2.20}$$

Set $D_n(t)=\frac{\rho_{2n}^q-\rho_{2n-2}^q}{h}(t)$ for $n=1,2,\dots,N$ and $0\le t<\infty$. Multiplying (2.20) by $\frac{\rho_{2n}^q-\rho_{2n-2}^q}{h}$, integrating the resulting equation over $[0,t]$, and summing the obtained equations from 1 to N, one can obtain

$$\frac{1}{2q}\sum D^2(t)h\le\frac{1}{2q}\sum D^2(0)h+\int_0^t\sum\left[\rho_{2n}^q\sum_{j=1}^{n}\dot{u}_{2j-1}-\rho_{2n-2}^q\sum_{j=1}^{n-1}\dot{u}_{2j-1}\right]D_n(s)hds. \tag{2.21}$$

The last term in (2.21) can be bounded by $C\int_0^t(\Sigma\dot{u}_{2n-1}^2h+1)\Sigma D_n^2(s)hds+C$ by summation by parts. Equation (2.19) follows then. □

The next lemma, which controls the expanding rate of interfaces and the total variation of velocity, can be deduced by the standard arguments (cf. [10], [11], [12], [13]). Then we omit the proof.

Lemma 2.6.

$$\begin{aligned} &c_3^{-1}\left\{\sum_{n=1}^{N-1}\rho_{2n}^{-1}(0)h+M^{-\gamma}\sum_{n=1}^{N-1}\rho_{2n}^{\gamma-1}(0)h[(1+c_4M^{\gamma}t)^{1/\gamma}-1]\right\}\\ &\leq\sum_{n=1}^{N-1}\rho_{2n}^{-1}(t)h\\ &\leq c_1^{-1}\left\{\sum_{n=1}^{N-1}\rho_{2n}^{-1}(0)h+c_2^{1/\gamma}t^{1/\gamma}\right\}, \end{aligned}\tag{2.22}$$

where $c_i(i=1,2,3,4)$ is the constant in (2.13), $M=\max_{[0,1]}\rho_0(y)$,

$$\sum_{n=1}^{N-1}|u_{2n+1}(t)-u_{2n-1}(t)|\leq C(1+t)^{1/\gamma},\tag{2.23}$$

$$\max_{1\leq n\leq N}|u_{2n-1}|\leq C(1+t)^{1/\gamma}\tag{2.24}$$

for $0\leq t<+\infty$.

To guarantee the convergence of the approximate solutions, we still need the higher order estimates and the Hölder continuous of $(\rho_{2n},u_{2n-1},\rho_{2n}\frac{u_{2n+1}-u_{2n-1}}{h})(t)$ with respect to t in $L^2[0,1]$-norm, which are given in the following two lemmas. The proof of these two lemmas is same as that in [12], based on the estimates we have obtained so far.

Lemma 2.7.

$$\sum\dot{u}_{2n-1}^2(t)h+\int_0^t\sum\rho_{2n}\left(\frac{\dot{u}_{2n+1}-\dot{u}_{2n-1}}{h}\right)^2hds\leq C,\tag{2.25}$$

$$\sum\frac{1}{h^2}\left(\rho_{2n}\frac{u_{2n+1}-u_{2n-1}}{h}-\rho_{2n-2}\frac{u_{2n-1}-u_{2n-3}}{h}\right)^2\cdot h\leq C(T),\tag{2.26}$$

$$\max_{1\leq n\leq N}\left|\rho_{2n}\frac{u_{2n+1}-u_{2n-1}}{h}\right|\leq C(T).\tag{2.27}$$

Lemma 2.8. *For any $t,\ s\in[0,T]$, $T>0$, it holds*

$$\sum|\rho_{2n}(t)-\rho_{2n}(s)|^2h\leq C|t-s|,\tag{2.28}$$

$$\sum\||u_{2n-1}(t)-u_{2n-1}(s)|^2h\leq C|t-s|,\tag{2.29}$$

$$\begin{aligned} &\sum\left|\rho_{2n}(t)\frac{u_{2n+1}(t)-u_{2n-1}(t)}{h}-\rho_{2n}(s)\frac{u_{2n+1}(s)-u_{2n-1}(s)}{h}\right|^2h\\ &\leq C|t-s|. \end{aligned}\tag{2.30}$$

With the desired estimates Lemmas 2.2–2.8 at hand, we are now in the position to prove Theorem 2.1 (cf. [8]). Define

$$\begin{aligned} \rho_h(t,y) &= \rho_{2n}(t), \\ u_h(t,y) &= \frac{[y-(n-\frac{1}{2})h]u_{2n+1} + [(n+\frac{1}{2})h - y]u_{2n-1}(t)}{h} \end{aligned}$$

for $(n-1/2)h < y < (n+1/2)h$. Then

$$\rho_h(u_h)_y(t,y) = \rho_{2n}(t)\frac{u_{2n+1}-u_{2n-1}}{h}.$$

Our main convergence result can be stated as follows.

THEOREM 2.9. *Suppose* ($A1'$)–($A3'$) *and* (2.1) *hold. There exist a subsequence of* $\{(\rho_h, u_h)\}$ *and* $\{\rho_h(u_h)_y\}$, *still labeled by* $\{(\rho_h, u_h)\}$ *and* $\{\rho_h(u_h)_y\}$ *for convenience, such that* $\{(\rho_h, u_h)\}$ *and* $\{\rho_h(u_h)_y\}$ *converge boundedly and almost everywhere on* $[0,T]\times[0,1]$ *for any* $T > 0$.

Proof. From our estimates, we know the functions (ρ_h^q, u_h) and $(\rho_h(u_h)_y)$, as the functions of y, have uniformly bounded total variations with respect to h for any fixed t. Let $t = t_m$ ($m = 1, 2, \ldots$) be a countable set which is everywhere dense on the segment $[0,T]$. By Helly's theorem and a diagonal process, from the family of the functions $\{(\rho_h, u_h)\}$ and $\{(\rho_h(u_h)_y\}$ one can select a sequence converging boundedly and almost everywhere in $y \in [0,1]$ on the dense set $t = t_m$ ($m = 1, 2, \ldots$) in $[0,T]$. Consequently, by Lebesgue's theorem, the subsequence also converges in L^2-norm on $t = t_m$ ($m = 1, 2, \ldots, N$). Next, with the help of Lemma 2.8, it is standard to show (ρ_h, u_h), $\rho_h(u_h)_y$ converge in $L^2[0,1]$ uniformly in $t \in [0,T]$. Then we can select a subsequence converge almost everywhere in $(y,t) \in [0,1]\times[0,T]$. □

Denote the limiting functions of $\{(\rho_h, u_h)\}$ and $\{\rho_h(u_h)_y\}$ by $(\rho, u)(y,t)$ and $\rho u_y(y,t)$, respectively. Then it is easy to verify (see [8]) that $(\rho, u)(y,t)$ is the weak solution of (2.1) and (2.2). The solution satisfies the corresponding limit version of the estimates in Lemmas 2.2–2.8. Part (3) in Theorem 2.1 is the limit version of (2.13) and the following equality:

$$\sum u_{2n-1}(t)h = \sum u_{2n-1}(0)h, \qquad 0 \le t \le \infty, \tag{2.31}$$

which is a consequence of $(2.8)_2$ and the boundary condition (2.9). This completes the proof of Theorem 2.1, which, in turn, proves Theorem 1.2.

3. Smoothness of the solution. In this section, we study the smoothness of the solutions constructed in section 2 and prove Theorem 1.5. At first, we have the following L^2-estimate of $\rho^{\beta/2}\rho_{yy}$, which is crucial to the improvement of regularity of the solution.

LEMMA 3.1. *Let* (ρ, u) *be the solution to* (2.1) *and* (2.2) *constructed in section* 2. *If the initial data satisfy*

$$\int_0^1 \rho^\beta \rho_{yy}^2(y,0)dy \le C \tag{3.1}$$

for $\beta \ge 4q-2$ *(*q *is the constant in* (A2$'$)*), then*

$$\int_0^1 \rho^\beta \rho_{yy}^2(y,t)dy \le C(T) \tag{3.2}$$

for $0 \le t \le T$.

Proof. For simplicity of presentation, we will only give the a priori estimate (3.2) for smooth solutions. The general version of Lemma 3.1 can be verified by a discrete analogue as in section 2.

At first, it follows from (2.1) that

$$\rho_{ty} = -\rho\rho_y u_y - \rho u_t - \rho P(\rho)_y.$$

Differentiating the above equation with respect to y, we get

$$\rho_{tyy} + \gamma A\rho^\gamma \rho_{yy} = -\rho_{yy}\rho u_y - \rho_y(u_t + (\rho u_y)_y) - \rho u_{ty} - \gamma^2 A\rho^{\gamma-1}\rho_y^2. \tag{3.3}$$

For the resulting equation by multiplying (3.3) by $\rho^\beta\rho_{yy}$, then integrating it over $[0,1]\times[0,t]$, we estimate each term on the right-hand side as follows.

First, (2.6) implies

$$\left|\int_0^t\int_0^1 \rho^\beta\rho_{yy}^2\rho u_y dyds\right| \le C(T)\int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 dyds. \tag{3.4}$$

Next, it follows from (2.5) that

$$\begin{aligned}&\left|\int_0^t\int_0^1 \rho^\beta\rho_{yy}\rho_y(u_t + (\rho u_y)_y)dyds\right|\\ &\le \int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 + \int_0^t \max\nolimits_{[0,1]}\rho^\beta\rho_y^2\int_0^1(u_t^2 + (\rho u_y)_y^2)ds\\ &\le \int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 dyds + C(T)\int_0^t \max\nolimits_{[0,1]}\rho^\beta\rho_y^2 ds,\end{aligned} \tag{3.5}$$

and

$$\begin{aligned}&\left|\int_0^t\int_0^1 \rho^\beta\rho_{yy}\rho u_{ty}dyds\right|\\ &\le \int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 dyds + C\int_0^t\int_0^1 \rho^{2+\beta}u_{ty}^2 dyds\\ &\le \int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 dyds + C,\end{aligned} \tag{3.6}$$

due to the fact that $2+\beta \ge 4q \ge 2$.

Finally, one has

$$\begin{aligned}&\left|\int_0^t\int_0^1 A\rho^{\gamma-1}\rho_y^2\rho^\beta\rho_{yy}dyds\right|\\ &\le \int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 dyds + \int_0^t\int_0^1 \rho^{2\gamma-2+\beta}\rho_y^4 dyds.\end{aligned} \tag{3.7}$$

Collecting estimates (3.4)–(3.7) yields

$$\begin{aligned}&\int_0^1 \rho^\beta\rho_{yy}^2(y,t)dy + \int_0^t\int_0^1 \rho^{\beta+\gamma}\rho_{yy}^2 dyds\\ &\le C(T)\int_0^t\int_0^1 \rho^\beta\rho_{yy}^2 dyds + C\int_0^t \max\nolimits_{[0,1]}\rho^\beta\rho_y^2 ds + \int_0^t\int_0^1 \rho^{2\gamma-2+\beta}\rho_y^4 dyds.\end{aligned} \tag{3.8}$$

The last two terms in (3.8) can be estimated as follows. First,

$$
\begin{aligned}
\max_{[0,1]}|\rho^{\beta/2}\rho_y| &\leq \int_0^1 |\rho^{\beta/2}\rho_y|dy + \int_0^1 |(\rho^{\beta/2}\rho_y)_y|dy \\
&\leq \left(\int_0^1 \rho^{\beta}\rho_y^2 dy\right)^{1/2} + C\int_0^1 \rho^{\frac{\beta}{2}-1}\rho_y^2 dy + \int_0^1 |\rho^{\beta/2}\rho_{yy}|dy. \qquad (3.9)
\end{aligned}
$$

Thus , if $\frac{\beta}{2}-1 \geq 2q-2$, i.e., $\beta \geq 4q-2$, then we get

$$
\max_{[0,1]}\rho^{\beta}\rho_y^2 \leq \int_0^1 \rho^{\beta}\rho_{yy}^2 dy + C(T). \qquad (3.10)
$$

We now turn to the last term in (3.8). In view of (2.6),

$$
\begin{aligned}
&\int_0^t\int_0^1 \rho^{2\gamma-2+\beta}\rho_y^4 dyds \\
&\leq \int_0^t \max_{[0,1]}\rho^{2\gamma+\beta-2q}\rho_y^2 \int_0^1 \rho^{2q-2}\rho_y^2 dyds \\
&\leq C(T)\int_0^t \max_{[0,1]}\rho^{2\gamma+\beta-2q}\rho_y^2 ds. \qquad (3.11)
\end{aligned}
$$

Since $2\gamma \geq 2q$, (3.10) and (3.11) imply

$$
\int_0^t\int_0^1 \rho^{2\gamma-2+\beta}\rho_y^4 dyds \leq C(T)\int_0^t\int_0^1 \rho^{\beta}\rho_{yy}^2 dyds + C(T). \qquad (3.12)
$$

It follows that

$$
\int_0^1 \rho^{\beta}\rho_{yy}^2(y,t)dy \leq C(T)\int_0^t\int_0^1 \rho^{\beta}\rho_{yy}^2(y,s)dyds + C(T).
$$

This, together with Gronwall's inequality, leads to the desired estimates. □

In the next lemma, we show that the assumption in Lemma 3.1 is true when the initial data satisfy some requirement in Eulerian coordinates.

LEMMA 3.2. *If* $\beta \geq \max\{2k+3+\alpha, 6k+1\}$, *then* $\int_0^1 \rho_0^{\beta}\rho_{0yy}^2(y)dy \leq C$ *provided* (A1) *holds and* $\int_a^b \rho_0^{\alpha}\rho_{0xx}^2(x)dx \leq C$.

Proof. It is easy to check the following relation between Eulerian coordinates and Lagrangian coordinates:

$$
\rho_{yy} = \rho^{-2}\rho_{xx} - \rho^{-3}\rho_x^2. \qquad (3.13)
$$

Thus

$$
\int_0^1 \rho_0^{\beta}\rho_{0yy}^2(y)dy \leq C\int_a^b (\rho_0^{\beta-3}\rho_{0xx}^2 + \rho_0^{\beta-5}\rho_{0x}^4)(x)dx. \qquad (3.14)
$$

On the other hand, (A2) implies

$$
\int_a^b \rho_0^{\beta-5}\rho_{0x}^4(x)dx \leq \max_{[a,b]}\rho_0^{\beta-2k-3}\rho_{0x}^2 \int_a^b \rho_0^{2k-2}\rho_{0x}^2(x)dx \leq C\max_{[a,b]}\rho_0^{\beta-2k-3}\rho_{0x}^2. \qquad (3.15)
$$

Using Sobolev's lemma and (A2), we obtain

$$\begin{aligned}
&\max_{[a,b]} |\rho_0^{\frac{\beta-2k-3}{2}} \rho_{0x}| \\
&\le C \int_a^b \left\{ \left| \rho_0^{\frac{\beta-2k-3}{2}} \rho_{0x} \right| + \rho_0^{\frac{\beta-2k-5}{2}} \rho_{0x}^2 + \left| \rho_0^{\frac{\beta-2k-3}{2}} \rho_{0xx} \right| \right\} dx \\
&\le C \int_a^b \left\{ \left| \rho_0^{\beta-2k-3} \rho_{0x}^2 \right| + \rho_0^{\frac{\beta-2k-5}{2}} \rho_{0x}^2 + \rho_0^{\beta-2k-3} \rho_{oxx}^2 dx \right\} dx + C. \qquad (3.16)
\end{aligned}$$

Lemma 3.2 follows from (3.14)–(3.16). □

We claim that the desired L^2-estimate of $\rho^{\alpha_1/2}\rho_{xx}$ in Eulerian coordinate can be bounded by the obtained estimate in Lagrangian coordinate by the following lemma.

LEMMA 3.3. *Let* (ρ, u) *be the solution stated in Theorem* 1.2 *to the free boundary problem* (1.4). *Then we have, for* $0 \le t \le T$,

$$\int_{a(t)}^{b(t)} \rho^{\alpha_1} \rho_{xx}^2(x,t) dx \le C(T) \left(\int_0^1 \rho^\beta \rho_{yy}^2(y,t) dy \right)^2 + C(T), \qquad (3.17)$$

if $\alpha_1 \ge \max\{\beta + 2k - 3, 6k - 2\}$.

Proof. It follows from (3.13) that

$$\begin{aligned}
&\int_{a(t)}^{b(t)} \rho^{\alpha_1} \rho_{xx}^2(x,t) dx \\
&= \int_0^1 \rho^{-1+\alpha_1} (\rho \rho_y^2 + \rho^2 \rho_{yy})^2 (y,t) dy \\
&\le C \int_0^1 (\rho^{1+\alpha_1} \rho_y^4 + \rho^{3+\alpha_1} \rho_{yy}^2)(y,t) dy. \qquad (3.18)
\end{aligned}$$

On the other hand, (2.6) implies

$$\int_0^1 \rho^{1+\alpha_1} \rho_y^4 dy \le \max_{[0,1]} \rho^{3-2q+\alpha_1} \rho_y^2 \int_0^1 \rho^{2q-2} \rho_y^2 dy \le C(T) \max_{[0,1]} \rho^{3-2q+\alpha_1} \rho_y^2, \qquad (3.19)$$

where $q = k + 1/2$.

Using Sobolev's lemma, (2.6), and the Cauchy inequality, we get

$$\begin{aligned}
&\max_{[0,1]} |\rho^{\frac{3-2q+\alpha_1}{2}} \rho_y| \\
&\le C \int_0^1 \left\{ \left| \rho^{\frac{3-2q+\alpha_1}{2}} \rho_y \right| + \rho^{\frac{1-2q+\alpha_1}{2}} \rho_y^2 + \left| \rho^{\frac{3-2q+\alpha_1}{2}} \rho_{yy} \right| \right\} dy \\
&\le C \int_0^1 (\rho^{4-2q+\alpha_1} \rho_y^2 + \rho^{4-2q+\alpha_1} \rho_{yy}^2) dy \\
&+ C \int_0^1 \rho^{-1} dy + C \int_0^1 \rho^{\frac{1-2q+\alpha_1}{2}} \rho_y^2 dy. \qquad (3.20)
\end{aligned}$$

Lemma 3.3 follows from (3.18)–(3.20). □

We now complete the proof of Theorem 1.5. To prove part (1) in Theorem 1.5, we only need to show the Hölder continuities described there. In fact, $(1.1)_2$ is uniformly parabolic with respect to u in the region $\{(x,t), a(t)+\delta \le x \le b(t)-\delta, 0 < t \le T\}$ for any $\delta > 0$ and $T > 0$ because $\rho > 0$ in that region. Thus, the standard parabolic theory (see [7] for instance) and the regularity of ρ and u which we have obtained imply the Hölder continuities indicated in part (1) of Theorem 1.5. Part (2) in Theorem 1.5 is the consequence of part (1), (1.11), and (1.13).

4. Uniqueness theorem. This section is devoted to the proof of Theorem 1.6. For this purpose, we prove the following Theorem 4.1 first. Theorem 1.6 follows from this theorem immediately.

Theorem 4.1. *Let* $(\rho_1, u_1)(y,t)$ *and* $(\rho_2, u_2)(y,t)$ *be two weak solutions to* (2.1) *and* (2.2) *in* $[0,1]\times[0,T]$ *with the properties corresponding to Definition* 1.1. *Then* $(\rho_1, u_1) = (\rho_2, u_2)$ *in* $[0,1]\times[0,T]$.

It should be noted that Theorem 4.1 particularly implies the following theorem.

Theorem 4.2. *The whole sequence of the approximate solutions* $\{(\rho_h, u_h)\}$ *constructed in section* 2 *converges as* $h \to 0$.

Proof of Theorem 4.1. Let (ρ_1, u_1) and (ρ_2, u_2) be the solutions stated in Theorem 4.1. We let

$$\varphi(y,t) = (\rho_1-\rho_2)(y,t),\ \psi(y,t) = \int_0^y (u_1-u_2)(z,t)dz \tag{4.1}$$

for $0 \le y \le 1$ and $0 \le t \le T$.

In the following, we may assume that (ρ_1, u_1) and (ρ_2, u_2) are suitably smooth since the following estimates are valid for the solutions with the regularity indicated in Theorem 2.1 by using the Friedrichs mollifier. In view of part (3) in Theorem 2.1 and the boundary condition $(2.2)_2$, we have

$$\varphi(0,t) = \varphi(1,t) = 0, \psi(0,t) = \psi(1,t) = 0 \tag{4.2}$$

for $0 \le t \le T$.

It follows from (2.1) and part (3) in Theorem 2.1 that

$$\begin{aligned} &\left(\frac{\varphi}{\rho_1\rho_2}\right)_t + \psi_{yy} = 0,\\ &\psi_t + (P(\rho_1)-P(\rho_2)) = \rho_1\psi_{yy} + \varphi u_{2y}. \end{aligned} \tag{4.3}$$

Multiplying $(4.3)_1$ by $\rho_2^{-1}\varphi$ and $(4.3)_2$ by $\rho_1^{-1}\psi$, and using the equations for (ρ_1, u_1) and (ρ_2, u_2), we get

$$\begin{aligned} &\left(\frac{1}{2}\rho_1^{-1}\psi^2\right)_t + \psi_y^2\\ &= \rho_1^{-1}\psi\varphi u_{2y} - \frac{1}{2}\psi^2 u_{1y} - (P(\rho_1)-P(\rho_2))\rho_1^{-1}\psi, \end{aligned} \tag{4.4}$$

$$\left(\frac{1}{2}\rho_1^{-1}\rho_2^{-2}\varphi^2\right)_t + \frac{1}{2}\rho_2^{-2}\varphi^2 u_{1y} + \rho_2^{-1}\varphi\psi_{yy} = 0. \tag{4.5}$$

We multiply $(4.3)_2$ by ψ_{yy} to get

$$\left(\frac{1}{2}\psi_y\right)^2 + \rho_1\psi_{yy}^2 = (P(\rho_1)-P(\rho_2))\psi_{yy} + \varphi u_{2y}\psi_{yy} + (\psi_t\psi_y)_y. \tag{4.6}$$

Integrating (4.4)–(4.6) over $[0,1]\times[0,t]$, using the boundary condition (4.2), and noting the regularity near the boundary for (ρ_1,u_1) and (ρ_2,u_2), we get, by virtue of the Cauchy inequality and (2.4)–(2.7), that

$$\begin{aligned}
&\frac{d}{dt}\int_0^1 \rho_1^{-1}\psi^2 dy+\int_0^1 \psi_y^2(y,t)dy\\
&\le C\left[\int_0^1 \rho_1^{-1}\psi^2(y,t)dy+\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(\rho_2u_{2y})^2(y,t)dy\right]\\
&+C\left[\int_0^1 \rho_1^{-1}\psi^2|\rho_1u_{1y}|(y,t)dy+\int_0^1 \rho_1^{-1}\varphi^2(y,t)dy\right]\\
&\le C(T)\left[\int_0^1 \rho_1^{-1}\psi^2(y,t)dy+\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(y,t)dy\right],
\end{aligned}\tag{4.7}$$

$$\begin{aligned}
&\frac{d}{dt}\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(y,t)dy\\
&\le C\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2|\rho_1u_{1y}|(y,t)dy+\frac{1}{4}\int_0^1 \rho_1\psi_{yy}^2(y,t)dy+C\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(y,t)dy\\
&\le C(T)\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(y,t)dy+\frac{1}{4}\int_0^1 \rho_1\psi_{yy}^2(y,t)dy,
\end{aligned}\tag{4.8}$$

and

$$\begin{aligned}
&\frac{d}{dt}\int_0^1 \frac{1}{2}\psi_y^2dy+\int_0^1 \rho_1\psi_{yy}^2(y,t)dy\\
&\le \frac{1}{4}\int_0^1 \rho_1\psi_{yy}^2(y,t)dy+C\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(y,t)dy+C\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(\rho_2u_{2y})^2(y,t)dy\\
&\le \frac{1}{4}\int_0^1 \rho_1\psi_{yy}^2(y,t)dy+C(T)\int_0^1 \rho_1^{-1}\rho_2^{-2}\varphi^2(y,t)dy
\end{aligned}\tag{4.9}$$

for $0\le t\le T$. Therefore,

$$\begin{aligned}
&\frac{d}{dt}\int_0^1 (\rho_1^{-1}\psi^2+\rho_1^{-1}\rho_2^{-2}\varphi^2+\psi_y^2)(y,t)dy\\
&+\int_0^1 (\psi_y^2+\rho_1\psi_{yy}^2)(y,t)dy\\
&\le C(T)\int_0^1 (\rho_1^{-1}\psi^2+\rho_1^{-1}\rho_2^{-2}\varphi^2)(y,t)dy
\end{aligned}\tag{4.10}$$

for $0\le t\le T$. Integrating (4.10) with respect to t, we get

$$\begin{aligned}
&\int_0^1 (\rho_1^{-1}\psi^2+\rho_1^{-1}\rho_2^{-2}\varphi^2)(y,t)dy\\
&\le C(T)\int_0^t\int_0^1 (\rho_1^{-1}\psi^2+\rho_1^{-1}\rho_2^{-2}\varphi^2)(y,s)dys.
\end{aligned}\tag{4.11}$$

Theorem 4.1 follows from (4.11) by applying Gronwall's inequality.

Theorem 4.2 is an immediate consequence of Theorem 4.1.

REFERENCES

[1] D. Hoff, *Global existence for* 1*-D compressible, isentropic Navier–Stokes equations in one space dimension with nonsmooth initial data*, Proc. Roy. Soc. Edinburgh Sect. A, 103 (1986), pp. 301–315.

[2] D. Hoff, *Strong convergence to global solutions for multidimensional flows of compressible, viscous fluids with polytropic equations of state and discontinuous initial data*, Arch. Rational Mech. Anal., 132 (1995), pp. 1–14.

[3] D. Hoff and D. Serre, *The failure of continuous dependence on initial data for the Navier–Stokes equations of compressible flow*, SIAM J. Appl. Math., 51 (1991), pp. 887–898.

[4] D. Hoff and T. P. Liu, *The inviscid limit for the Navier–Stokes equations of compressible isentropic flow with shock data*, Indiana Univ. Math. J., 38 (1989), pp. 861–915.

[5] S. Kawashima, *Systems of a Hyperbolic-Parabolic Composite Type, with Applications to the Equations of Magnetohydrodynamics*, Ph.D. thesis, Kyoto University, Kyoto, Japan, 1983.

[6] S. Kawashima and T. Nishida, *Global solutions to the initial value problem for the equations of one-dimensional motion of viscous polytropic gases*, J. Math. Kyoto Univ., 21 (1981), pp. 825–837.

[7] O. A. Ladyzenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[8] T. P. Liu, Z. Xin, and T. Yang, *Vacuum states of compressible flow*, Discrete Contin. Dynam. Systems, 4 (1998), pp. 1–32.

[9] T. Makino, *On a local existence theorem for the evolution equations of gaseous stars*, in Patterns and Wave-Qualitative Analysis of Nonlinear Differential Equations, T. Nishida, M. Mimura, and H. Fujii, eds., North–Holland, Amsterdam, 1986, pp. 459–479.

[10] S. Matusu-Necasova, M. Okada, and T. Makino, *Free boundary value problems for the equation of one-dimensional motion of viscous gas* (II), Japan J. Indust. Appl. Math., 12 (1995), pp. 195–203.

[11] T. Nishida, *Equations of fluid dynamics–free surface problems*, Comm. Pure Appl. Math., 39 (1986), pp. 221–238.

[12] M. Okada, *Free boundary value problems for the equation of one-dimensional motion of viscous gas*, Japan J. Appl. Math., 6 (1989), pp. 161–177.

[13] M. Okada and T. Makino, *Free boundary problem for the equations of spherically symmetrical motion of viscous gas*, Japan J. Indust. Appl. Math., 10 (1993), pp. 219–235.

[14] D. Serre, *Solutions faibles globales des equations de Navier–Stokes pour un fluide compressible*, C. R. Acad. Sci. Paris Sér. 1 Math., 303 (1986), pp. 639–642.

[15] D. Serre, *Sur l'equation mondimensionnelle d'un fluide visqueux, compressible et conducteur de chaleur*, C. R. Acad. Sci. Paris Sér. 1 Math., 303 (1986), pp. 703–706.

[16] Z. Xin, *Blow-up of smooth solutions to the compressible Navier–Stokes equations with compact density*, Comm. Pure Appl. Math., 51 (1998), pp. 229–240.

[17] Z. Xin, *Zero dissipation limit to rarefaction waves for one-dimensional Navier–Stokes equations for compressible isentropic gases*, Comm. Pure Appl. Math., Vol. 46 (1993), pp. 621–665.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1192–1213

GROUP ANALYSIS OF DIFFERENTIAL EQUATIONS AND GENERALIZED FUNCTIONS[*]

MICHAEL KUNZINGER[†] AND MICHAEL OBERGUGGENBERGER[‡]

Abstract. We present an extension of the methods of classical Lie group analysis of differential equations to equations involving generalized functions (in particular: distributions). A suitable framework for such a generalization is provided by Colombeau's theory of algebras of generalized functions. We show that under some mild conditions on the differential equations, symmetries of classical solutions remain symmetries for generalized solutions. Moreover, we introduce a generalization of the infinitesimal methods of group analysis that allows us to compute symmetries of linear and nonlinear differential equations containing generalized function terms. Thereby, the group generators and group actions may be given by generalized functions themselves.

Key words. algebras of generalized functions, Lie symmetries of differential equations, group analysis, delta waves, Colombeau algebras

AMS subject classifications. 46F30, 35D05, 35D10, 35A30, 58G35

PII. S003614109833450X

1. Introduction.

Symmetry properties of distributions and group invariant distributional solutions (in particular: fundamental solutions) to particular types of linear differential operators have been studied by Methée [22], Tengstrand [36], Szmydt and Ziemian [33], [34], [35], [38]. A systematic investigation of the transfer of classical group analysis of differential equations into a distributional setting is due to Berest and Ibragimov [2], [3], [4], [5], [18], again with a view to determining fundamental solutions of certain linear partial differential equations. A survey of the lastnamed studies including a comprehensive bibliography can be found in the third volume of [19]. As these approaches use methods from classical distribution theory, their range is confined to linear equations and linear transformations of the dependent variables.

Algebras of generalized functions offer the possibility of going beyond these limitations towards a generalization of group analysis to genuinely nonlinear problems involving singular terms, like distributions or discontinuous nonlinearities. In the present paper we develop a theory of group analysis of differential equations in algebras of generalized functions that allows a satisfactory treatment of such problems. This line of research has been initiated in [28] and has been taken up in [21]. Applications to different types of algebras of generalized functions can be found in [30] and [31].

The plan of the paper is as follows. Section 2 provides a short introduction to the theory of algebras of generalized functions in the sense of Colombeau. In section 3 we consider systems of partial differential equations together with a classical symmetry group G that transforms smooth solutions into smooth solutions. Assuming polynomial bounds on the action of G, we can extend it to generalized functions belonging to Colombeau algebras and ask whether G remains a symmetry group for generalized solutions. In section 3.1 we develop methods based on a factorization

[*] Received by the editors February 24, 1998; accepted for publication (in revised form) September 23, 1999; published electronically May 11, 2000.

http://www.siam.org/journals/sima/31-6/33450.html

[†] Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria (Michael.Kunzinger@univie.ac.at). This author was supported by research grant P-10472 MAT of the Austrian Science Foundation.

[‡] Institut für Technische Mathematik, Geometrie und Bauinformatik, Universität Innsbruck, Technikerstr.13, A-6020 Innsbruck, Austria (michael@mat1.uibk.ac.at).

property of the transformed system of equations. Essentially, polynomial bounds on the factors suffice to give a positive answer. In the scalar case we show this to be automatically satisfied whenever the equation contains at least one of the derivatives of the solution as an isolated term. While the conditions of section 3.1 concern some mild assumptions on the algebraic structure of the equations, section 3.2 develops a topological criterion, applicable to systems of linear equations: the existence of a $\mathcal{C}^\infty$-continuous homogeneous right inverse guarantees a positive answer as well. Along the way we give examples of nonlinear symmetry transformations of shock and delta wave solutions to linear and nonlinear systems.

The purpose of section 4 is to develop the general theory, allowing the equations and the group action (hence also its generators) to be given by generalized functions. Using the characterization of Colombeau generalized functions by their generalized pointvalues established in [27] as well as results on Colombeau solutions to ordinary differential equations (ODEs), we show that the classical procedure for computing symmetries can literally be transferred to the generalized function situation. The defining equations are derived as usual, but their solutions are sought in generalized functions. This enlarges the reservoir of possible symmetries of classical equations and allows the study of symmetries of equations with singular terms. An example is provided by a conservation law with nondifferentiable flux function.

The remainder of the introduction is devoted to fixing notations and recalling some basic definitions from group analysis of differential equations. We basically follow the notations and terminology of [29]. Thus for the action of a Lie group G on some manifold M, assumed to be an open subset of some space $\mathcal{X} \times \mathcal{U}$ of independent and dependent variables (with $\dim(\mathcal{X}) = p$ and $\dim(\mathcal{U}) = q$), we write $g \cdot (x, u) = (\Xi_g(x, u), \Phi_g(x, u))$. Transformation groups are always supposed to act regularly on M. If Ξ_g does not depend on u, the group action is called projectable. Elements of the Lie algebra $\mathfrak{g}$ of G as well as the corresponding vector fields on M will typically be denoted by $\mathbf{v}$ and the one-parameter subgroup generated by $\mathbf{v}$ by $\eta \to \exp(\eta\mathbf{v})$. $M^{(n)}$ denotes the n-jet space of M; the nth prolongation of a group action g or vector field $\mathbf{v}$ is written as $\mathrm{pr}^{(n)}g$ or $\mathrm{pr}^{(n)}\mathbf{v}$, respectively. Any system S of nth order differential equations in p dependent and q independent variables can be written in the form

$$\Delta_\nu(x, u^{(n)}) = 0, \qquad 1 \le \nu \le l,$$

where the map

$$\begin{gathered}\Delta : \mathcal{X} \times \mathcal{U}^{(n)} \to \mathbb{R}^l, \\ (x, u^{(n)}) \to (\Delta_1(x, u^{(n)}), \dots, \Delta_l(x, u^{(n)}))\end{gathered}$$

is supposed to be smooth. Hence S is identified with the subvariety

$$S_\Delta = \{(x, u^{(n)}) : \Delta(x, u^{(n)}) = 0\}$$

of $\mathcal{X} \times \mathcal{U}^{(n)}$. For any $f : \Omega \subseteq \mathcal{X} \to \mathcal{U}$, Γ_f is the graph of f and $\Gamma_f^{(n)} := \{(x, \mathrm{pr}^{(n)} f(x)) : x \in \Omega\}$ is the graph of the n-jet of f.

2. Colombeau algebras. Already at a very early stage of development of the theory of distributions it became clear that it is impossible to embed the space $\mathcal{D}'(\Omega)$ of distributions over some open subset Ω of $\mathbb{R}^n$ into an associative, commutative algebra $(\mathcal{A}(\Omega), +, \circ)$ satisfying

(i) $\mathcal{D}'(\Omega)$ is linearly embedded into $\mathcal{A}(\Omega)$ and $f(x) \equiv 1$ is the unity in $\mathcal{A}(\Omega)$.

(ii) There exist derivation operators $\partial_i : \mathcal{A}(\Omega) \to \mathcal{A}(\Omega)$ $(i = 1, \dots, n)$ that are linear and satisfy the Leibnitz rule.

(iii) $\partial_i|_{\mathcal{D}'(\Omega)}$ is the usual partial derivative $(i = 1, \dots, n)$.

(iv) $\circ|_{\mathcal{C}(\Omega)\times\mathcal{C}(\Omega)}$ coincides with the pointwise product of functions.

This is the well-known impossibility result of Schwartz [32]. Replacing $\mathcal{C}(\Omega)$ by $\mathcal{C}^{(k)}(\Omega)$ does not alter this result. On the other hand, many problems involving differentiation and nonlinearities in the presence of singular objects require a method of coping with this situation in a consistent manner (cf., e.g., [24], [26], [37]). By the above, the best possible result would consist in constructing an algebra $\mathcal{A}(\Omega)$ satisfying (i)–(iii) and

(iv′) $\circ|_{\mathcal{C}^\infty(\Omega)\times\mathcal{C}^\infty(\Omega)}$ coincides with the pointwise product of functions.

The actual construction of algebras enjoying these optimal properties is due to Colombeau [8], [9] (see also [1], [24]). The basic idea underlying his theory (in its simplest—the so-called special—form) is that of embedding the space of distributions into a factor algebra of $\mathcal{C}^\infty(\Omega)^I$ $(I = (0, 1])$ via regularization by convolution with a fixed "mollifier" $\rho \in \mathcal{S}(\mathbb{R}^n)$ with $\int \rho(x)\,dx = 1$. In order to motivate the definition below let $\rho_\varepsilon(x) := \varepsilon^{-n}\rho(\frac{x}{\varepsilon})$ and let $u \in \mathcal{E}'(\mathbb{R}^n)$ (the space of compactly supported distributions on $\mathbb{R}^n$). The sequence $(u * \rho_\varepsilon)_{\varepsilon\in I}$ converges to u in $\mathcal{D}'(\mathbb{R}^n)$. Taking this sequence as a representative of u we obtain an embedding of $\mathcal{D}'(\mathbb{R}^n)$ into the algebra $\mathcal{C}^\infty(\mathbb{R})^I$. However, embedding $\mathcal{C}^\infty(\mathbb{R}^n) \subseteq \mathcal{D}'(\mathbb{R}^n)$ into this algebra via convolution as above will not yield a subalgebra since of course $(f * \rho_\varepsilon)(g * \rho_\varepsilon) \neq (fg) * \rho_\varepsilon$ in general. The idea, therefore, is to factor out an ideal $\mathcal{N}(\mathbb{R}^n)$ such that this difference vanishes in the resulting quotient. In order to construct $\mathcal{N}(\mathbb{R}^n)$ it is obviously sufficient to find an ideal containing all differences $(f*\rho_\varepsilon)_{\varepsilon\in I} - (f)_{\varepsilon\in I}$. Taylor expansion of $f*\rho_\varepsilon - f$ shows that this term will vanish faster than any power of ε (uniformly on compact sets, in all derivatives), provided we additionally suppose that $\int \rho(x)x^\alpha\,dx = 0\,\forall\,\alpha \in \mathbb{N}_0^n$ with $|\alpha| \geq 1$. The set of all such sequences is not an ideal in $\mathcal{C}^\infty(\mathbb{R}^n)^I$, so we shall replace $\mathcal{C}^\infty(\mathbb{R}^n)^I$ by the set of *moderate* sequences $\mathcal{E}_M(\mathbb{R}^n)$ whose every derivative is bounded uniformly on compact sets by some inverse power of ε.

Thus we define the Colombeau algebra $\mathcal{G}(\Omega)$ as the quotient algebra $\mathcal{E}_M(\Omega)/\mathcal{N}(\Omega)$, where

$$\begin{aligned}\mathcal{E}_M(\Omega) := \{&(u_\varepsilon)_{\varepsilon\in I} \in \mathcal{C}^\infty(\Omega)^I : \forall K \subset\subset \Omega\ \forall \alpha \in \mathbb{N}_o^n\ \exists p \in \mathbb{N} \text{ with}\\ &\sup_{x\in K} |\partial^\alpha u_\varepsilon(x)| = O(\varepsilon^{-p}) \text{ as } \varepsilon \to 0\},\\ \mathcal{N}(\Omega) := \{&(u_\varepsilon)_{\varepsilon\in I} \in \mathcal{C}^\infty(\Omega)^I : \forall K \subset\subset \Omega\ \forall \alpha \in \mathbb{N}_o^n\ \forall q \in \mathbb{N},\\ &\sup_{x\in K} |\partial^\alpha u_\varepsilon(x)| = O(\varepsilon^{q}) \text{ as } \varepsilon \to 0\}.\end{aligned}$$

Equivalence classes of sequences $(u_\varepsilon)_{\varepsilon\in I}$ in $\mathcal{G}(\Omega)$ will be denoted by $\mathrm{cl}[(u_\varepsilon)_{\varepsilon\in I}]$. $\mathcal{G}(\Omega)$ is a differential algebra containing $\mathcal{E}'(\Omega)$ as a linear subspace via the embedding $\iota : u \to \mathrm{cl}[(u * \rho_\varepsilon)_{\varepsilon\in I}]$ depending on a mollifier $\rho \in \mathcal{S}(\mathbb{R}^n)$ as above. ι commutes with partial derivatives and coincides with $u \to \mathrm{cl}[(u)_{\varepsilon\in I}]$ on $\mathcal{D}(\Omega)$, thus rendering it a faithful subalgebra of $\mathcal{G}(\Omega)$. The functor $\Omega \to \mathcal{G}(\Omega)$ is a fine sheaf of differential algebras on $\mathbb{R}^n$ and there is a unique sheaf morphism $\hat{\iota}$ extending the above embedding to $\mathcal{C}^\infty(\,.\,) \hookrightarrow \mathcal{D}'(\,.\,) \hookrightarrow \mathcal{G}(\,.\,)$. $\hat{\iota}$ commutes with partial derivatives, and its restriction to $\mathcal{C}^\infty$ is a sheaf morphism of algebras.

We shall also consider the algebra $\mathcal{G}_\tau(\Omega) = \mathcal{E}_\tau(\Omega)/\mathcal{N}_\tau(\Omega)$ of tempered generalized functions, where

$$\mathcal{O}_M(\Omega) = \{f \in \mathcal{C}^\infty(\Omega) : \forall \alpha \in \mathbb{N}_o^n\ \exists p > 0\ \sup_{x\in\Omega}(1 + |x|)^{-p}|\partial^\alpha f(x)| < \infty\},$$

$$\begin{aligned}\mathcal{E}_\tau(\Omega) = \{&(u_\varepsilon)_{\varepsilon\in I} \in (\mathcal{O}_M(\Omega))^I : \forall\alpha \in \mathbb{N}_o^n\ \exists p > 0,\\ &\sup_{x\in\Omega}(1+|x|)^{-p}|\partial^\alpha u_\varepsilon(x)| = O(\varepsilon^{-p})\ (\varepsilon \to 0)\},\\ \mathcal{N}_\tau(\Omega) = \{&(u_\varepsilon)_{\varepsilon\in I} \in (\mathcal{O}_M(\Omega))^I : \forall\alpha \in \mathbb{N}_o^n\ \exists p > 0\ \forall\ q > 0,\\ &\sup_{x\in\Omega}(1+|x|)^{-p}|\partial^\alpha u_\varepsilon(x)| = O(\varepsilon^{q})\ (\varepsilon \to 0)\}.\end{aligned}$$

The map ι defined above is a linear embedding of $\mathcal{S}'(\mathbb{R}^n)$ into $\mathcal{G}_\tau(\mathbb{R}^n)$ commuting with partial derivatives and making

$$\mathcal{O}_C(\mathbb{R}^n) = \{f \in \mathcal{C}^\infty(\mathbb{R}^n) : \exists p > 0\ \forall\alpha \in \mathbb{N}_o^n\ \sup_{x\in\mathbb{R}^n}(1+|x|)^{-p}|\partial^\alpha f(x)| < \infty\}$$

a faithful subalgebra. Elements of $\mathcal{O}_M(\Omega)$ are called *slowly increasing*. Componentwise insertion of elements of $\mathcal{G}$ into slowly increasing functions yields well-defined elements of $\mathcal{G}$. Thus, in $\mathcal{G}$ not only polynomial combinations of distributions (e.g., δ^2) make sense but also expressions like $\sin(\delta)$ have a well-defined meaning. The importance of $\mathcal{G}_\tau(\Omega)$ for our purposes stems from the fact that elements of this algebra can even be composed with each other (again by componentwise insertion, cf. [16], [20]), a necessary prerequisite for generalizing symmetry methods; see section 4. Especially in the theory of ODEs in the generalized function context it is often useful to consider the algebra $\widetilde{\mathcal{G}}_\tau(\Omega \times \Omega')$ whose elements satisfy $\mathcal{G}$-bounds in the Ω-variables and $\mathcal{G}_\tau$-bounds in the Ω'-variables (cf. [16] or [20]). Elements of Colombeau algebras are usually denoted by capital letters with the understanding that $(u_\varepsilon)_{\varepsilon\in I}$ denotes an arbitrary representative of $U \in \mathcal{G}$.

Nonlinear operations with distributions in $\mathcal{G}(\Omega)$ depend not only on the distributions themselves but also on the regularization procedure used in the embedding process. Thus the difference of two representatives $(u_\varepsilon)_{\varepsilon\in I}$, $(v_\varepsilon)_{\varepsilon\in I}$ of generalized functions U, resp. V, may have $\mathcal{D}'$-limit 0 as $\varepsilon \to 0$ without U and V being equal in $\mathcal{G}(\Omega)$. Nevertheless U and V are to be considered "equal in the sense of distributions" or *associated* with each other ($U \approx V$). Moreover, U is called associated with some distribution w if $u_\varepsilon \to w$ in $\mathcal{D}'$. If such a w exists (which need not be the case, cf. δ^2), it is to be seen as the distributional "shadow" of U. For example, all powers of the Heaviside function are associated with each other without being equal in the algebra itself. Also, $x\delta = 0$ in $\mathcal{D}'(\mathbb{R})$, so $x\delta \approx 0$ in $\mathcal{G}(\mathbb{R})$, but $x\delta \neq 0$ in $\mathcal{G}(\mathbb{R})$. These examples illustrate a general principle: assigning nonlinear properties to elements of the vector space $\mathcal{D}'(\Omega)$ amounts to introducing additional information which is reflected in a more rigid concept of equality within $\mathcal{G}(\Omega)$ compared to that in $\mathcal{D}'(\Omega)$. This strict concept of equality allows for much more refined ways of infinitesimal modelling. On the $\mathcal{D}'$-level (the level of association) this additional information is lost in the limit-process $\varepsilon \to 0$.

Generalized numbers (i.e., the ring of constants in case Ω is connected) in any of the above algebras will be denoted by $\mathcal{R}$. Componentwise insertion of points into representatives of generalized functions yields well defined elements of $\mathcal{R}$.

We note that there exist variants of Colombeau algebras that allow a canonical embedding of distributions (independent of a fixed mollifier as above). The basic idea for constructing these algebras is to replace the index set I by the space of *all* possible mollifiers. Our choice of the special variants of Colombeau algebras is aimed at notational simplicity. However, all results presented in what follows carry over to the respective full variants of the algebras. Moreover, recently there have been introduced global versions of Colombeau algebras, defined intrinsically on manifolds and displaying the analogues of (i)–(iv) (with ∂_i replaced by Lie-derivatives with respect to smooth vector fields); see [14]. For applications of the theory to nonlinear partial differential equations (PDEs) see [24] and the literature cited therein; for applications to mathematical physics and numerics, cf. [6], [10], and [37].

3. Transfer of classical symmetry groups.

3.1. Factorization properties. The first question to be answered in trying to extend the applicability of classical group analysis to generalized solutions concerns permanence properties of classical symmetries: Let G be the symmetry group of some system S of PDEs and consider S within the framework of $\mathcal{G}(\Omega)$. Under which conditions do elements of G also transform generalized solutions into other generalized solutions? It is the aim of this and the following section to answer this question. To begin with, let us fix some terminology.

Definition 3.1. *Let G be a projectable local group of transformations acting on some open set $\mathcal{M} \subseteq \mathcal{X} \times \mathcal{U}$ according to $g \cdot (x, u) = (\Xi_g(x), \Phi_g(x, u))$. g is called slowly increasing if the map $u \to \Phi_g(x, u)$ is slowly increasing, uniformly for x in compact sets. g is strictly slowly increasing if $\Phi_g \in \mathcal{O}_M(\mathcal{M})$. If $\Omega \subseteq \mathcal{X}$, $U \in \mathcal{G}(\Omega)$ and g is (strictly) slowly increasing, the action of g on U is defined as the element*

$$gU := \mathrm{cl}[((\Phi_g \circ (id \times u_\varepsilon)) \circ \Xi_g^{-1})_{\varepsilon \in I}] \tag{3.1}$$

of $\mathcal{G}(\Xi_g(\Omega))$.

If U is a smooth function, (3.1) reproduces the classical notion of group action on functions. Henceforth we make the tacit assumption that the differential equations under consideration are of a form that allows for an insertion of elements of Colombeau generalized functions (i.e., the function Δ representing the equations on the prolongation space is slowly increasing). Also, slowly increasing group actions are always understood to be projectable. Analogous to the classical setting we give the following definition.

Definition 3.2. *Let S be some system of differential equations with p variables and q unknown functions. A solution of S in $\mathcal{G}$ is an element $U \in (\mathcal{G}(\Omega))^q$, with $\Omega \subseteq \mathcal{X}$ open, which solves the system with equality in $(\mathcal{G}(\Omega))^l$. A symmetry group of S in $\mathcal{G}$ is a local transformation group acting on $\mathcal{X} \times \mathcal{U}$ such that if U is a solution of the system in $\mathcal{G}$, $g \in G$ and $g \cdot U$ is defined, then also $g \cdot U$ is a solution of S in $\mathcal{G}$.*

Let us take a look at the transition problem from classical to generalized symmetry groups on the level of representatives. Thus, let G be a slowly increasing symmetry group of some differential equation

$$\Delta(x, u^{(n)}) = 0. \tag{3.2}$$

This means that if f is a classical solution, i.e., if $\Delta(x, \mathrm{pr}^{(n)} f(x)) = 0\ \forall x$ then also $\Delta(x, \mathrm{pr}^{(n)}(g \cdot f)(x)) = 0$. Now let $U \in \mathcal{G}(\Omega)$ be a generalized solution to (3.2). Then for any representative $(u_\varepsilon)_{\varepsilon \in I}$ of U there exists some $(n_\varepsilon)_{\varepsilon \in I} \in \mathcal{N}(\Omega)$ such that $\forall x$ and $\forall\ \varepsilon$ we have

$$\Delta(x, \mathrm{pr}^{(n)} u_\varepsilon(x)) = n_\varepsilon(x). \tag{3.3}$$

In particular, the differential equation (3.2) need not be satisfied for even one single value of ε. This basic observation displays quite fundamental obstacles to a direct utilization of the classical symmetry group properties of G in order to obtain statements on the status of G in the Colombeau-setting. Therefore we have to derive properties of symmetry groups that are better suited to allow a transfer to differential algebras. The starting point for our considerations is a slight modification of a well known factorization property of smooth maps (cf. [29, Proposition 2.10]) as follows.

PROPOSITION 3.3. *Let F be a smooth mapping from some manifold M to $\mathbb{R}^k$ $(k \leq n = \dim(M))$, let $f : (-\eta_o, \eta_o) \times M \to \mathbb{R}$ be smooth and suppose that $f(\eta, .)$ vanishes on the zero set $\mathcal{S}_F$ of F, identically in η. If F is of maximal* rank *$(= k)$ on $\mathcal{S}_F$ then there exist smooth functions $Q_1, \dots, Q_k : (-\eta_o, \eta_o) \times M \to \mathbb{R}$ such that*

$$f(\eta, m) = Q_1(\eta, m)F_1(m) + \dots + Q_k(\eta, m)F_k(m)$$

$\forall(\eta, m) \in (-\eta_o, \eta_o) \times M$. □

We are mainly interested in the following application of Proposition 3.3.

THEOREM 3.4. *Let*

$$\Delta_\nu(x, u^{(n)}) = 0, \qquad 1 \leq \nu \leq l \tag{3.4}$$

be a nondegenerate system of PDEs. Let $G = \{g_\eta : \eta \in (-\eta_o, \eta_o)\}$ be a one parameter symmetry group of (3.4) *and set $g_\eta \cdot (x, u) = (\Xi_\eta(x, u), \Phi_\eta(x, u))$. Then there exist $\mathcal{C}^\infty$-functions $Q_{\mu\nu} : (- \eta_o, \eta_o) \times \mathcal{V} \to \mathbb{R}$ $(1 \leq \mu, \nu \leq l$, $\mathcal{V}$ an open subset of $\mathcal{M}^{(n)})$ such that if $u : \Omega \subseteq \mathbb{R}^p \to \mathbb{R}^q$ is smooth and $g_\eta u$ exists we have*

$$\begin{aligned} &\Delta_\nu(\Xi_\eta(x, u(x)), \mathrm{pr}^{(n)}(g_\eta u)(\Xi_\eta(x, u(x)))) \\ &= \sum_{\mu=1}^{l} Q_{\mu\nu}(\eta, x, \mathrm{pr}^{(n)}u(x))\Delta_\mu(x, \mathrm{pr}^{(n)}u(x)) \end{aligned} \tag{3.5}$$

on the domain of $g_\eta u$ for $1 \leq \nu \leq l$.

Proof. Denote by z the coordinates on $\mathcal{M}^{(n)}$. That g_η is an element of the symmetry group of the system is equivalent with

$$\Delta(z) = 0 \Rightarrow \Delta_\nu(\mathrm{pr}^{(n)}g_\eta(z)) = 0 \quad (1 \leq \nu \leq l)$$

$\forall \eta$ and z such that this is defined. Δ is of maximal rank because (3.4) is nondegenerate. Hence, by Proposition 3.3 there exist $\mathcal{C}^\infty$-functions $Q_{\mu\nu} : (-\eta_o, \eta_o) \times \mathcal{V} \to \mathbb{R}$ $(1 \leq \mu \leq l$, $\mathcal{V}$ an open subset of $\mathcal{M}^{(n)})$ such that

$$\Delta_\nu(\mathrm{pr}^{(n)}g_\eta(z)) = \sum_{\mu=1}^{l} Q_{\mu\nu}(\eta, z)\Delta_\mu(z). \tag{3.6}$$

Now for a smooth function $u : \Omega \subseteq \mathbb{R}^p \to \mathbb{R}^q$ as in our assumption and $x \in \Omega$ we set

$$z_u(x) := (x, \mathrm{pr}^{(n)}u(x)) \in \mathcal{M}^{(n)}. \tag{3.7}$$

Then by definition $\mathrm{pr}^{(n)}g_\eta(z_u(x)) = (\Xi_\eta(x, u(x)), \mathrm{pr}^{(n)}(g_\eta u)(\Xi_\eta(x, u(x))))$, so the result follows. □

For a single PDE $\Delta(x, \mathrm{pr}^{(n)}u) = 0$, (3.5) takes the simpler form

$$\Delta(\Xi_\eta(x, u(x)), \mathrm{pr}^{(n)}(g_\eta u)(\Xi_\eta(x, u(x)))) = Q(\eta, x, \mathrm{pr}^{(n)}u(x))\Delta(x, \mathrm{pr}^{(n)}u(x)). \tag{3.8}$$

Theorem 3.4 will be one of our main tools in transferring classical symmetry groups of (systems of) PDEs into the setting of algebras of generalized functions.

PROPOSITION 3.5. *Let $\eta \to g_\eta$ be a slowly increasing one parameter symmetry group of* (3.4)*. If $P_{\mu\nu} := (Q_{\mu\nu}(\eta, \Xi_{-\eta}(\,.\,), \mathrm{pr}^{(n)} u_\varepsilon(\Xi_{-\eta}(\,.\,))))_{\varepsilon\in I}$ belongs to $\mathcal{E}_M(\Omega)$ for $1 \le \mu, \nu \le l$ and every $(u_\varepsilon)_{\varepsilon\in I} \in \mathcal{E}_M(\Omega)$, then $\eta \to g_\eta$ is a symmetry group of* (3.4) *in $\mathcal{G}$ as well. This condition is satisfied if*

$$(x, u^{(n)}) \to Q_{\mu\nu}(\eta, x, u^{(n)})$$

is slowly increasing in the $u^{(n)}$-variables, uniformly in x on compact sets for $1 \le \mu, \nu \le l$ and every η.

Proof. It suffices to observe that (3.5) gives

$$\begin{aligned}&\Delta_\nu(x, \mathrm{pr}^{(n)}(g_\eta u)(x))\\ &= \sum_{\mu=1}^{l} Q_{\mu\nu}(\eta, \Xi_{-\eta}(x), \mathrm{pr}^{(n)} u(\Xi_{-\eta}(x)))\Delta_\mu(\Xi_{-\eta}(x), \mathrm{pr}^{(n)} u(\Xi_{-\eta}(x))).\end{aligned}$$

For any solution $U \in \mathcal{G}(\Omega)$ with representative $u = (u_\varepsilon)_{\varepsilon\in I}$, this expression is in $\mathcal{N}(\Omega)$ since $P_{\mu\nu} \in \mathcal{E}_M(\Omega)$ for each μ, ν, and every $\Delta_\mu(\Xi_{-\eta}(\,.\,), \mathrm{pr}^{(n)} u(\Xi_{-\eta}(\,.\,)))$ is in $\mathcal{N}(\Omega)$ because U is a solution and $\Xi_{-\eta}$ is a diffeomorphism. □

Example 3.6. The system

$$\begin{aligned} U_t + UU_x &= 0,\\ V_t + UV_x &= 0,\\ U\,|_{\{t=0\}} = U_o, \quad V\,|_{\{t=0\}} &= V_o, \end{aligned} \tag{3.9}$$

may serve as a simplified model for a one-dimensional, elastic material of high density in a nearly plastic state. It was analyzed in [25], where solutions $U, V \in \mathcal{G}_{s,g}(\mathbb{R} \times [0,\infty))$, $U_o, V_o \in \mathcal{G}_{s,g}(\mathbb{R})$ were constructed and studied ($\mathcal{G}_{s,g}$ is a variant of the Colombeau algebra with global instead of local bounds). In the following we present some applications of the above results to this system (for a more detailed study, see [21]). For $U_o' \ge 0$ (3.9) has a unique solution (U, V) in $\mathcal{G}_{s,g}(\mathbb{R} \times [0,\infty))$ with $\partial_x U \ge 0$. We consider solutions in $\mathcal{G}_{s,g}(\mathbb{R} \times [0,\infty))$ with initial data $U_o(x) = u_L + (u_R - u_L)H(x)$ and $V_o(x) = v_L + (v_R - v_L)H(x)$, where H is a generalized Heaviside function with $H' \ge 0$, i.e., H is a member of $\mathcal{G}_{s,g}(\mathbb{R})$ with a representative $(h_\varepsilon)_{\varepsilon\in I}$ coinciding with the classical Heaviside function Y off the interval $[-\varepsilon, \varepsilon]$. For $u_L < u_R$ the solution (U, V) is associated with the rarefaction wave

$$u(x,t) = \begin{cases} u_L, & x \le u_L t,\\ \frac{x}{t}, & u_L t \le x \le u_R t,\\ u_R, & u_R t \le x, \end{cases} \tag{3.10}$$

$$v(x,t) = \begin{cases} v_L, & x \le u_L t,\\ \left(\frac{v_R - v_L}{u_R - u_L}\right)\frac{x}{t} + \left(\frac{v_L u_R - v_R u_L}{u_R - u_L}\right), & u_L t \le x \le u_R t,\\ v_R, & u_R t \le x. \end{cases} \tag{3.11}$$

However, choosing different generalized Heaviside functions for modelling the initial data U_o, respectively V_o, we may obtain a superposition of the rarefaction wave (3.10) in u with a shock wave

$$v(x,t) = v_L + (v_R - v_L)Y(x - ct) \tag{3.12}$$

with arbitrary shock speed c, $u_L \leq c \leq u_R$. We are going to construct a one parameter symmetry $\eta \to g_\eta$ of (3.9) which transforms any of the solutions (3.11), (3.12) into a shock wave solution as $\eta \to \pm\infty$. For this we employ the two-dimensional Lorentz-transformation $(\eta, (x,t)) \to (x\cosh(\eta) - t\sinh(\eta), -x\sinh(\eta) + t\cosh(\eta))$ with infinitesimal generator $X_o = -t\partial_x - x\partial_t$. Then $X := X_o + (u^2-1)\partial_u$ generates a projectable one-parameter symmetry group of (3.9). Assuming that $-1 < u_{\mathrm{L}} < u_{\mathrm{R}} < 1$, we can extend the solution (U,V) to the region $\Omega = \mathbb{R}^2 \setminus \{(x,t) : t \leq 0, u_{\mathrm{R}}t \leq x \leq u_{\mathrm{L}}t\}$ by the method of characteristics applied to representatives. Then the Lorentz-transformed solutions

$$\begin{aligned} \tilde{u}_\varepsilon(x,t) &= -\tanh(\eta - \operatorname{Artanh}(u_\varepsilon(x\cosh(\eta) + t\sinh(\eta),\\ &\qquad x\sinh(\eta) + t\cosh(\eta)))), \end{aligned} \tag{3.13}$$

$$\tilde{v}_\varepsilon(x,t) = v_\varepsilon(x\cosh(\eta) + t\sinh(\eta), x\sinh(\eta) + t\cosh(\eta)) \tag{3.14}$$

(with Artanh the inverse of tanh) are well defined at least on $\mathbb{R} \times (0,\infty)$. The factorization property (3.5) in this case reads

$$\begin{aligned} &(\partial_t \tilde{u}_\varepsilon + \tilde{u}_\varepsilon \partial_x \tilde{u}_\varepsilon)(x,t)\\ &\quad = \big((\partial_t u_\varepsilon + u_\varepsilon \partial_x u_\varepsilon)/(\cosh^3(\operatorname{Artanh}(u_\varepsilon - \eta))\cosh(\operatorname{Artanh}(u_\varepsilon)))\big)\big(\Xi_\eta^{-1}(x,t)\big) \end{aligned} \tag{3.15}$$

and similarly for the second line in (3.9), demonstrating that $(\widetilde{U}, \widetilde{V})$ is again a solution. For each η, $\widetilde{U}$ is associated with a piecewise smooth function which converges to ∓ 1 as $\eta \to \pm\infty$. Observing that the coordinate transformations in (3.13), (3.14) approach boosts in the directions $(\mp 1, 1)$ as $\eta \to \pm\infty$, we see that the functions associated with $\widetilde{V}$ converge to the shock wave $v_L + (v_R - v_L)Y(x \pm t)$ as $\eta \to \pm\infty$, for whatever solution V given in (3.11) or (3.12).

Although Proposition 3.5 provides a manageable algorithm to determine if classical symmetry groups carry over to generalized solutions it would certainly be preferable to have criteria at hand that allows us to judge directly from the given PDE if the factors $P_{\mu\nu}$ behave nicely (given slowly increasing group actions). The first step in this direction is gaining control over the behavior of the map $z \to \mathrm{pr}^{(n)} g_\eta(z)$, defined on $\mathcal{M}^{(n)}$.

Proposition 3.7. *If $\eta \to g_\eta$ is a (strictly) slowly increasing group action on $\mathcal{M}$ then $z \to \mathrm{pr}^{(n)} g_\eta(z)$ is (strictly) slowly increasing as well.*

Proof. Let $N := \dim(\mathcal{M}^{(n)})$. For $z = (z_1, \ldots, z_p, z_{p+1}, \ldots, z_q, \ldots, z_N) \in \mathcal{M}^{(n)}$ we choose some smooth function $h : \mathcal{X} \to \mathcal{U}$ satisfying $z = z_h(z_1, \ldots, z_p)$, with $z_h(x)$ as in (3.7). Then we set $x := (z_1, \ldots, z_p)$, $u = (z_{p+1}, \ldots, z_q)$, $\widetilde{x} = \Xi_\eta(x)$ and $\widetilde{u} = \Phi_\eta(x,u)$. By the definition of prolonged group actions we have to find estimates for every

$$A_s := ((\Phi_\eta \circ (id \times h)) \circ \Xi_{-\eta})^{(s)}\,(\widetilde{x}) \tag{3.16}$$

(where (s) denotes the derivative of order s) in terms of z. The above formula contains the components of $\mathrm{pr}^{(n)} g(z)$ of order s ($s \leq n$). Note that the particular choice of h has no influence on (3.16), i.e., A_s depends exclusively on z. To compute A_s explicitly we use the formula for higher order derivatives of composite functions (see [11]). Denoting by Υ_m the group of permutations of $\{1, \ldots, m\}$ we have

$$A_s(r_1, \ldots, r_s) = \sum_{i=1}^{s} \sum_{\substack{k \in \mathbb{N}^i\\ |k|=s}} \sum_{\sigma \in \Upsilon_s} \frac{1}{i!k!} (\Phi_\eta \circ (id \times h))^{(i)}((\widetilde{x}))(t_1, \ldots, t_i), \tag{3.17}$$

where

$$t_1 = \Xi_{-\eta}^{(k_1)}(\widetilde{x})(r_{\sigma(1)},\dots,r_{\sigma(k_1)}),\dots,t_i = \Xi_{-\eta}^{(k_i)}(\widetilde{x})(r_{\sigma(s-k_i+1)},\dots,r_{\sigma(s)})$$

and

$$(3.18)\ (((\Phi_\eta \circ (id \times h)))^{(i)}(x)(t_1,\dots,t_i) = \sum_{j=1}^{i} \sum_{\substack{l\in\mathbb{N}^j \\ |l|=i}} \sum_{\tau\in\Upsilon_i} \frac{1}{j!l!}\Phi_\eta^{(j)}(x,u)(s_1,\dots,s_j),$$

where

$$s_1 = (id\times h)^{(l_1)}(x)(t_{\tau(1)},\dots,t_{\tau(l_1)}),\ \dots\ ,s_j = (id\times h)^{(l_j)}(x)(t_{\tau(i-l_j+1)},\dots,t_{\tau(i)}).$$

Each s_m consists of sums of products of certain $t_{\tau(k)}$ with certain z_l and an analogous assertion holds for the $\Phi_\eta^{(j)}(x,u)(s_1,\dots,s_j)$. Hence from (3.17) and (3.18) the result follows. □

Returning to our original task of finding a priori estimates for the factors $P_{\mu\nu}$, even with the aid of Proposition 3.7 we still need some information about the explicit form of the $Q_{\mu\nu}$ to go on. In general this seems quite difficult to achieve. However, there is a large and important class of PDEs that allow a priori statements on the concrete form of the factorization. Namely, we are going to show that each scalar PDE in which at least u or one of its derivatives appears as a single term with constant coefficient belongs to this class.

Consider a scalar PDE $\Delta(x,u^{(n)})=0$ together with a symmetry group $\eta\to g_\eta$. Then we have

$$\Delta(z)=0 \quad\Rightarrow\quad \Delta(\mathrm{pr}^{(n)}g_\eta(z))=0.$$

Set $F(z) := \Delta(z)$, $f(z) := \Delta(\mathrm{pr}^{(n)}g_\eta(z))$ and $N=\dim(\mathcal{M}^{(n)})$. Suppose that in a neighborhood of some $\bar{z}$ with $F(\bar{z})=0$ we have $\frac{\partial F}{\partial z_k}>0$ for some $1\le k\le N$. Then by the implicit function theorem, locally there exists a smooth function $\psi:\mathbb{R}^{N-1}\to\mathbb{R}$ such that in a suitable neighborhood of $\bar{z}$ we have

$$F(z)=0 \quad\Leftrightarrow\quad z_k=\psi(z'),$$

where $z'=(z_1,\dots,\hat{z_k},\dots,z_N)$ (meaning that the component z_k is missing from z'). It follows that

$$F(z) = (z_k-\psi(z'))\int_0^1 \frac{\partial F}{\partial z_k}(z_1,\dots,z_{k-1},\tau z_k+(1-\tau)\psi(z'),\dots,z_N)\,d\tau,$$

and on the other hand

$$f(z) = (z_k-\psi(z'))\int_0^1 \frac{\partial f}{\partial z_k}(z_1,\dots,z_{k-1},\tau z_k+(1-\tau)\psi(z'),\dots,z_N)\,d\tau.$$

Thus in the said neighborhood we have

$$(3.19)\qquad f(z) = F(z)\frac{\int_0^1 \frac{\partial f}{\partial z_k}(z_1,\dots,z_{k-1},\tau z_k+(1-\tau)\psi(z'),\dots,z_N)d\tau}{\int_0^1 \frac{\partial F}{\partial z_k}(z_1,\dots,z_{k-1},\tau z_k+(1-\tau)\psi(z'),\dots,z_N)d\tau},$$

provided the denominator of this expression is $\neq 0$. In particular, if for some constant $c \neq 0$ we have $\frac{\partial F}{\partial z_k} \equiv c$ in a neighborhood of $\bar{z}$ then (3.19) simplifies to

$$f(z) = \frac{1}{c} F(z) \int_0^1 \frac{\partial f}{\partial z_k}(z_1, \ldots, z_{k-1}, \tau z_k + (1-\tau)\psi(z'), \ldots, z_N) d\tau. \tag{3.20}$$

After these preparations we can state the following theorem.

Theorem 3.8. *Let $\eta \to g_\eta$ be a slowly increasing symmetry group of the equation $\Delta(x, u^{(n)}) = 0$. Set $N = \dim(\mathcal{M}^{(n)})$ and suppose that $\frac{\partial \Delta}{\partial z_k} \equiv c \neq 0$ for some $p+1 \leq k \leq N$. Then $\eta \to g_\eta$ is a symmetry group of $\Delta(x, u^{(n)}) = 0$ in $\mathcal{G}$.*

Proof. Without loss of generality we may assume $c = 1$. Using the above notations we have $F(z) = z_k - \psi(z')$, so (3.20) implies

$$f(z) = F(z) \int\limits_0^1 \frac{\partial f}{\partial z_k}(z_1, \ldots, z_{k-1}, \tau z_k + (1-\tau)(z_k - F(z)), \ldots, z_N) d\tau =: F(z) Q(\eta, z).$$

From Proposition 3.7 we know that $z \to f(z)$ is slowly increasing in the $u^{(n)}$-variables (i.e., in those z_i with $i > p$), uniformly in $x = (z_1, \ldots, z_p)$ on compact sets. Since F is slowly increasing we infer that $Q(\eta, z_u(x)) \in \mathcal{E}_M(\Omega)$ for any $u \in \mathcal{E}_M(\Omega)$ (with z_u as in (3.7)). Finally,

$$\Delta(x, \mathrm{pr}^{(n)}(g_\eta u)(x)) = \Delta(\Xi_{-\eta}(x), \mathrm{pr}^{(n)} u(\Xi_{-\eta}(x))) Q(\eta, \Xi_{-\eta}(x), \mathrm{pr}^{(n)} u(\Xi_{-\eta}(x))).$$

Since $\Xi_{-\eta}$ is a diffeomorphism, it follows that if $U = \mathrm{cl}[u]$ solves the equation, so does $g_\eta U$. □

As the proof shows, we can drop the assumption $p + 1 \leq k$ if we require the group action to be strictly slowly increasing. It is clear that many PDEs satisfy the requirements of Theorem 3.8. For example, in the Hopf equation $\Delta(x, t, u, u_x, u_t) = u_t + u u_x$ or $\Delta(z_1, \ldots, z_5) = z_5 + z_3 z_4$ one can take $k = 5$. Note, however, that not every symmetry group of this equation is automatically slowly increasing. Theorem 3.8 constitutes a useful tool for transferring classical symmetry groups to Colombeau algebras.

Example 3.9. We consider the initial value problem for the nonlinear transport equation

$$\begin{aligned} &U_t + \lambda \cdot \nabla_x U = f(U), \\ &U\,|_{\{t=0\}} = U_o \end{aligned} \tag{3.21}$$

with $t \in \mathbb{R}, x, \lambda \in \mathbb{R}^n$. It has unique solutions in $\mathcal{G}(\mathbb{R}^{n+1})$, given $U_o \in \mathcal{G}(\mathbb{R}^n)$, provided $f \in \mathcal{O}_M$ is globally Lipschitz (see [24]). If in addition f is bounded and the initial data are distributions with discrete support, say $U_0(x) = \sum_{i,j} a_{ij} \delta^{(i)}(x - \xi_j)$ with $\xi_j \in \mathbb{R}^n, i \in \mathbb{N}_0^n$, then the generalized solution is associated with the delta wave $v + w$ where

$$v(x, t) = \sum_{i,j} a_{ij} \delta^{(i)}(x - \lambda t - \xi_j) \tag{3.22}$$

and w is the smooth solution to $w_t + \lambda \cdot \nabla_x w = f(w)$, $w(0) = 0$.

The vector field $X = c f(u) \partial_u$ generates an infinitesimal symmetry of (3.21) for arbitrary $c \in \mathbb{R}$. With $F(u) := \int du/f(u)$, the corresponding Lie point transformation is

$$(x, t, u) \to (\widetilde{x}, \widetilde{t}, \widetilde{u}) = (x, t, F^{-1}(c\eta + F(u))). \tag{3.23}$$

This provides a well-defined nonlinear transformation of the generalized solution $U \in \mathcal{G}(\mathbb{R}^{n+1})$, provided that the right hand side in (3.23) is slowly increasing.

In the example

$$U_t + \lambda \cdot \nabla_x U = \tanh(U) \tag{3.24}$$

the generalized solution is associated with $v(x,t)$ and w vanishes identically. Applying (3.23) we obtain (due to Theorem 3.8) the new generalized solution

$$\widetilde{U}(x,t) = \operatorname{Arsinh}\left(e^{c\eta}\sinh(U(x,t))\right) \tag{3.25}$$

(with Arsinh the inverse of sinh). We are going to show that $\widetilde{U}$ is still associated with the delta wave v in (3.22). To simplify the argument we assume $n = 1, \lambda = 0$ and $U_0(x) = \delta^{(i)}(x)$. Representatives of U, resp. $\widetilde{U}$, are $u_\varepsilon(x,t) = \operatorname{Arsinh}(e^t \sinh(\rho_\varepsilon^{(i)}(x)))$ and $\widetilde{u}_\varepsilon(x,t) = \operatorname{Arsinh}(e^{c\eta+t}\sinh(\rho_\varepsilon^{(i)}(x)))$. For $\psi \in \mathcal{D}(\mathbb{R}^2)$ we have

$$\begin{aligned} I_\varepsilon^i &:= \int\int \widetilde{u}_\varepsilon(x,t)\psi(x,t)dxdt \\ &= \int\int\int_0^1 \theta(e^{c\eta+t}, \sigma\varepsilon^{-i-1}\rho^{(i)}(x))d\sigma\varepsilon^{-i}\rho^{(i)}(x)\psi(\varepsilon x,t)dxdt, \end{aligned}$$

where $\theta(\alpha, y) := \frac{d}{dy}\operatorname{Arsinh}(\alpha\sinh(y))$ for $\alpha > 0$, $y \in \mathbb{R}$. Since θ is bounded by $\max(1,\alpha)$ and $\lim_{|y|\to\infty}\theta(\alpha,y) = 1$ it follows that $I_\varepsilon^0 \to \int \psi(0,t)dt$, so $\widetilde{U}$ is associated with the delta function on the t-axis, as desired. For $i \geq 1$ we write

$$\begin{aligned} I_\varepsilon^i = \int\int\int_0^1 &(\theta(e^{c\eta+t}, \sigma\varepsilon^{-i-1}\rho^{(i)}(x)) - 1)d\sigma\varepsilon^{-i}\rho^{(i)}(x)\psi(\varepsilon x,t)dxdt \\ &+ (-1)^i \int\int \rho(x)\partial_x^i\psi(\varepsilon x,t)dxdt. \end{aligned}$$

Here the second term converges to $(-1)^i \int \partial_x^i \psi(0,t)$ and the first term goes to zero since $\int_0^1 |\theta(\alpha,\sigma y) - 1|d\sigma \leq \frac{2|\alpha^2-1|}{\alpha|y|}(1 - e^{-|y|})$ for $y \neq 0$. This proves the claim for $\rho \in \mathcal{D}(\mathbb{R})$. For $\rho \in \mathcal{S}(\mathbb{R})$ splitting the x-integral into one from $-\frac{1}{\sqrt{\varepsilon}}$ to $\frac{1}{\sqrt{\varepsilon}}$ and one over $|x| \geq \frac{1}{\sqrt{\varepsilon}}$ gives the same result.

3.2. Continuity properties. In this section we work out a different strategy for transferring classical point symmetries into the $\mathcal{G}$-setting. This approach, suggested in [28], consists in a more topological way of looking at the transfer problem by using continuity properties of differential operators. As we have pointed out in the discussion following (3.3), the main obstacle against directly applying classical symmetry groups componentwise to representatives of generalized solutions is that the differential equations need not be satisfied componentwise. However, there are certain classes of partial differential operators that do allow such a direct application. Consider a linear partial differential operator P giving rise to an equation

$$PU = 0 \tag{3.26}$$

in $\mathcal{G}$ and let G be a classical slowly increasing symmetry group of (3.26). Furthermore, suppose that P possesses a continuous homogeneous (but not necessarily linear) right inverse Q. If $U = \mathrm{cl}[u]$ is a solution to (3.26) in $\mathcal{G}(\Omega)$ then there exists some $n \in \mathcal{N}(\Omega)$ such that

$$Pu = n.$$

Since Q is a right inverse of P this implies

$$P(u_\varepsilon - Qn_\varepsilon) = 0 \quad \forall \varepsilon \in I. \tag{3.27}$$

Also, $Qn \in \mathcal{N}(\Omega)$ due to the continuity and homogeneity assumption on Q. If $g \in G$, (3.27) implies

$$P(g(u_\varepsilon - Qn_\varepsilon)) = 0 \quad \forall \varepsilon \in I.$$

By definition,

$$P(gU) = \mathrm{cl}[P(gu)] = \mathrm{cl}[P(g(u - Qn))],$$

so gU is a solution as well. Summing up, G is a symmetry group in $\mathcal{G}$. The following result will serve to secure the existence of a right inverse as above for a large class of linear differential operators.

Proposition 3.10. *Let E, F be Fréchet spaces and A a continuous linear map from E onto F. Then A has a continuous homogeneous right inverse $B : F \to E$.*

Proof. See [23, p. 364]. □

From these preparations we conclude the following theorem.

Theorem 3.11. *Let*

$$\Delta_\nu(x, u^{(n)}) = 0, \qquad \nu = 1, \dots, l$$

be a system of linear PDEs with slowly increasing Δ_ν and let $\eta \to g_\eta$ be a slowly increasing symmetry group of this system. Assume that the operator defined by the left hand side is surjective $(\mathcal{C}^\infty(\Omega))^l \to (\mathcal{C}^\infty(\Omega))^l$. Then $\eta \to g_\eta$ is a symmetry group for the system in $\mathcal{G}(\Omega)$ as well. □

The assumptions of Theorem 3.11 are automatically satisfied for any linear partial differential operator with constant coefficients on an arbitrary convex open domain (see [17, section 10.6]).

Example 3.12. The system of one-dimensional linear acoustics

$$\begin{aligned} P_t + U_x &= 0, \\ U_t + P_x &= 0 \end{aligned} \tag{3.28}$$

is transformed via $U = V - W, P = V + W$ into

$$\begin{aligned} V_t \,+\, V_x &= 0, \\ W_t - W_x &= 0. \end{aligned} \tag{3.29}$$

Using the infinitesimal generators $\Phi(v)\partial_v + \Psi(w)\partial_w$ (Φ, Ψ arbitrary smooth functions) of (3.29) we obtain symmetry transformations for (3.28) of the form

$$\widetilde{U} = F^{-1}\left(\eta + F\left(\frac{1}{2}(P+U)\right)\right) - G^{-1}\left(\theta + G\left(\frac{1}{2}(P-U)\right)\right)$$

$$\widetilde{P} = F^{-1}\left(\eta + F\left(\frac{1}{2}(P+U)\right)\right) + G^{-1}\left(\theta + G\left(\frac{1}{2}(P-U)\right)\right)$$

with arbitrary diffeomorphisms F, G. Since (3.28) satisfies the assumptions of Theorem 3.11 on $\Omega = \mathbb{R}^2$ it follows that any slowly increasing transformation of this form is a symmetry of (3.28). In particular, this includes nonlinear transformations of distributional solutions; cf. Example 3.13.

In the remainder of this section we discuss the interplay between symmetry groups and solutions of PDEs in the sense of association. Consider

$$\Delta_\nu(x, u^{(n)}) \approx 0, \qquad 1 \le \nu \le l \tag{3.30}$$

in $\mathcal{G}$. A slowly increasing symmetry group of the corresponding system

$$\Delta(x, u^{(n)}) = 0, \qquad 1 \le \nu \le l$$

is called a symmetry group in the sense of association if it transforms solutions of (3.30) into other such solutions. The first question to be answered in this context is whether one can derive conditions on the form of the factorization (3.8) that will yield symmetry groups in the sense of association. It is clear that a sufficient condition is to suppose that Q depends exclusively on η and x. Distributional solutions to linear PDEs arise as a special case of (3.30) and have been treated in [4]. There, the validity of (3.8) with Q depending on η and x only is actually used to *define* symmetry groups in $\mathcal{D}'$. In order to remain within the classical distributional framework, the admissible group transformations in [4] are restricted to projectable ones acting linearly in the dependent variables. On the other hand, the method developed there is even applicable to linear equations containing distributional terms which allows one to use invariance methods to compute fundamental solutions.

Second, if u is a solution to $\Delta(x, u^{(n)}) = 0$ in $\mathcal{G}(\Omega)$ possessing an associated distribution, one may ask for which group actions g this implies that gu as well possesses an associated distribution. This is certainly the case for admissible transformations in the above sense. On the other hand, we have already seen in Example 3.9 that even genuinely nonlinear symmetry transformations may preserve association properties.

The next example shows that nonlinear group actions may transform distributional solutions in Examples 3.9 and 3.12 into more complicated distributional solutions or into generalized solutions in $\mathcal{G}(\mathbb{R}^2)$ not admitting associated distributions.

Example 3.13. We consider the equation $U_t + \lambda U_x = 0$ arising in (3.21) with $n = 1$ or in (3.29). We have already observed that $\widetilde{U} = F^{-1}(\eta + F(U))$ defines a symmetry transformation for arbitrary diffeomorphisms F. Here we take $F \in \mathcal{C}^\infty(\mathbb{R})$, $F' > 0$, $F(y) = \operatorname{sign}(y)\sqrt{|y|}$ for $|y| \ge 1$. We wish to compute $\widetilde{U}$ when $U \in \mathcal{G}(\mathbb{R}^2)$ is a delta wave solution $U(x,t) \approx \delta^{(i)}(x - \lambda t)$. We take U as the class of $\rho_\varepsilon^{(i)}(x - \lambda t)$ with $\rho \in \mathcal{D}([-1,1])$. We have when $\eta \ge 0$:

(i) If $i = 0$, that is, $U \approx \delta(x - \lambda t)$, then $\widetilde{U} \approx F^{-1}(\eta + F(0)) + \delta(x - \lambda t)$;

(ii) If $i = 1$, that is, $U \approx \delta'(x - \lambda t)$, then
$\widetilde{U} \approx F^{-1}(\eta + F(0)) + 2\eta \int \sqrt{|\rho'(y)|}\, dy\ \delta(x - \lambda t) + \delta'(x - \lambda t)$;

(iii) If $i \ge 2$ then $\widetilde{U}$ does not admit an associated distribution.

To see this, we may assume that $\lambda = 0$ and write $a_\varepsilon(x) := \rho_\varepsilon^{(i)}(x)$ for brevity. Note that $F^{-1}(y) = \operatorname{sign}(y)y^2$ for $|y| \ge 1$. Let $A_\varepsilon = \{x \in [-\varepsilon, \varepsilon] : |a_\varepsilon(x)| \ge (\eta + 1)^2\}$. If $x \in A_\varepsilon$ and $a_\varepsilon(x) \ge 0$ then $\eta + F(a_\varepsilon(x)) \ge 1$ and $F^{-1}(\eta + F(a_\varepsilon(x))) = \eta^2 + 2\eta\sqrt{a_\varepsilon(x)} + a_\varepsilon(x)$. Also, if $x \in A_\varepsilon$ and $a_\varepsilon(x) < 0$ then $\eta + F(a_\varepsilon(x)) \le -1$ and $F^{-1}(\eta + F(a_\varepsilon(x))) = -\eta^2 + 2\eta\sqrt{|a_\varepsilon(x)|} + a_\varepsilon(x)$. The functions $F^{-1}(\eta + F(a_\varepsilon))$, $|a_\varepsilon(x)|$ and $\sqrt{|a_\varepsilon(x)|}$ are bounded on the complement of A_ε. Thus

$$\begin{aligned}
&\int\int_{-\varepsilon}^{\varepsilon} F^{-1}(\eta + F(a_\varepsilon(x)))\psi(x,t)dxdt \\
&= \int\int_{A_\varepsilon} \Big(\pm\, \eta^2 + 2\eta\sqrt{|a_\varepsilon(x)|} + a_\varepsilon(x)\Big)\psi(x,t)dxdt + O(\varepsilon) \\
&= \int\int_{-\varepsilon}^{\varepsilon} \Big(2\eta\sqrt{|a_\varepsilon(x)|} + a_\varepsilon(x)\Big)\psi(x,t)dxdt + O(\varepsilon)
\end{aligned}$$

while

$$\int\int_{|x|\geq\varepsilon} F^{-1}(\eta + F(a_\varepsilon(x)))\psi(x,t)\, dxdt \to F^{-1}(\eta + F(0)) \int\int \psi(x,t)\, dxdt.$$

It follows that $F^{-1}(\eta + F(a_\varepsilon(x)))$ converges in $\mathcal{D}'(\mathbb{R}^2)$ if and only if $2\eta\sqrt{|a_\varepsilon|} + a_\varepsilon$ admits an associated distribution. A simple computation yields the particular results (i), (ii), (iii).

4. Generalized group actions.

Although the methods introduced in the previous sections enable an application of large classes of classical symmetry groups to elements of Colombeau algebras, they are but the first step in a theory of generalized group analysis of differential equations. In this section we develop an extension of the methods of group analysis that will allow to consider symmetry groups of differential equations whose actions are generalized functions themselves.

4.1. Generalized transformation groups.

Simple examples indicate the necessity of extending the methods of group analysis of PDEs to equations involving generalized functions themselves.

Example 4.1. Considering (3.21) in $\mathcal{G}_\tau$ with a *generalized* function $f = \mathrm{cl}[(f_\varepsilon)_{\varepsilon\in I}] \in \mathcal{G}_\tau$ we can apply the classical algorithm for calculating symmetry groups componentwise to the equation

$$\partial_t u_\varepsilon + \lambda \cdot \nabla_x u_\varepsilon = f_\varepsilon(u_\varepsilon),$$

thereby obtaining infinitesimal generators with generalized coefficient functions. Thus the question arises in which sense such generators induce symmetries of the differential equation. More generally, one can consider differential equations in $\mathcal{G}_\tau$ of the form

$$P(x, U^{(n)}) = 0,$$

where P is a generalized function.

As is indicated by Example 4.1, composition of generalized functions will inevitably occur in a generalization of group analysis. For this purpose, we shall apply suitable variants of Colombeau algebras for the following considerations, namely $\mathcal{G}_\tau(\mathbb{R}^n)$ and $\widetilde{\mathcal{G}}_\tau(\mathbb{R}\times\mathbb{R}^n) = \widetilde{\mathcal{G}}_\tau(\mathbb{R}^{1+n})$.

Definition 4.2. *A generalized group action on* $\mathbb{R}^n$ *is an element* Φ *of* $(\widetilde{\mathcal{G}}_\tau(\mathbb{R}^{1+n}))^n$ *such that*

(i) $\Phi(0, \,.\,) = \mathrm{id}$ in $(\mathcal{G}_\tau(\mathbb{R}^n))^n$;

(ii) $\Phi(\eta_1 + \eta_2, \,.\,) = \Phi(\eta_1, \Phi(\eta_2, \,.\,))$ in $(\widetilde{\mathcal{G}}_\tau(\mathbb{R}^{2+n}))^n$.

Before we turn to an infinitesimal description of generalized group actions let us shortly recall some basic definitions from [27] that are needed for a pointvalue characterization of generalized functions which in turn plays a fundamental role in the following considerations. Thus for any open set $\Omega \subseteq \mathbb{R}^n$ we set

$$\Omega_M := \{(x_\varepsilon)_{\varepsilon\in I} \in \Omega^I : \exists p > 0 \ \exists\ \eta > 0 \ |x_\varepsilon| \le \varepsilon^{-p} \ (0 < \varepsilon < \eta)\}.$$

On Ω_M we define an equivalence relation by

$$(x_\varepsilon)_{\varepsilon\in I} \sim (y_\varepsilon)_{\varepsilon\in I} \ \Leftrightarrow\ \forall q > 0 \ \exists \eta > 0 \ |x_\varepsilon - y_\varepsilon| \le \varepsilon^q \ (0 < \varepsilon < \eta)$$

and set $\widetilde{\Omega} := \Omega_M/\sim$. $\widetilde{\Omega}$ is called the set of generalized points corresponding to Ω. The set of compactly supported points is defined as

$$\widetilde{\Omega}_c = \{\widetilde{x} \in \widetilde{\Omega} : \exists \text{ representative } (x_\varepsilon)_{\varepsilon\in I} \ \exists K \subset\subset \Omega \ \exists \eta > 0 \ :\ x_\varepsilon \in K,\ \varepsilon \in (0, \eta)\}.$$

Note that for $\Omega = \mathbb{R}$ we have $\widetilde{\Omega} = \mathcal{R}$. Theorems 2.4, 2.7, and 2.10 of [27] establish that elements of $\mathcal{G}(\Omega)$, $\widetilde{\mathcal{G}}_\tau(\Omega)$ or $\widetilde{\mathcal{G}}_\tau(\Omega\times\Omega')$ are uniquely determined by their pointvalues in $\widetilde{\Omega}_c$, $\widetilde{\Omega}$, or $\widetilde{\Omega}_c \times \widetilde{\Omega}'$, respectively. For the theory of ODEs in the Colombeau framework we refer to [16].

Definition 4.3. *Let $\xi = (\xi_1, \dots, \xi_n) \in (\mathcal{G}_\tau(\mathbb{R}^n))^n$. The generalized vector field $X = \sum_{i=1}^n \xi_i(x)\partial_{x_i}$ is called $\mathcal{G}$-complete if the initial value problem*

$$\begin{aligned} \dot{x}(t) &= \xi(x(t)), \\ x(t_o) &= \widetilde{x}_o \end{aligned}$$

is uniquely solvable in $\mathcal{G}(\mathbb{R})^n$ for any $\widetilde{x}_o \in \mathcal{R}^n$ and any $t_o \in \mathbb{R}$.

Definition 4.4. *Let Φ be a generalized group action on $\mathbb{R}^n$ and set*

$$\xi := \frac{d}{d\eta}\Big|_0 \Phi(\eta, \,.\,) \in (\mathcal{G}_\tau(\mathbb{R}^n))^n.$$

If the generalized vector field $X = \sum_{i=1}^n \xi_i(x)\partial_{x_i}$ is $\mathcal{G}$-complete, then X is called the infinitesimal generator of Φ. In this case, Φ is also called $\mathcal{G}$-complete.

By [16], every generalized vector field with $\mathcal{G}_\tau$-components whose gradient is of L^∞-log-type is $\mathcal{G}$-complete. The notion of infinitesimal generator is well-defined due to the following proposition.

Proposition 4.5. *Every $\mathcal{G}$-complete generalized group action is uniquely determined by its infinitesimal generator.*

Proof. Let Φ', Φ'' be two $\mathcal{G}$-complete generalized group actions with the same infinitesimal generator $X = \sum_{i=1}^n \xi_i(x)\partial_{x_i}$. Then both functions satisfy

$$\frac{d}{d\eta}\Phi(\eta, x) = \frac{d}{d\mu}\Big|_0 \Phi(\eta + \mu, x) = \frac{d}{d\mu}\Big|_0 \Phi(\mu, \Phi(\eta, x)) = \xi(\Phi(\eta, x)).$$

Now given any $\widetilde{x} \in \mathcal{R}^n$, it follows that both $\eta \to \Phi'(\eta, \widetilde{x})$ and $\eta \to \Phi''(\eta, \widetilde{x})$ solve the initial value problem

$$\begin{aligned} \dot{x}(\eta) &= \xi(x(\eta)), \\ x(0) &= \widetilde{x}. \end{aligned}$$

By assumption this entails that $\Phi'(\,.\,, \widetilde{x}) = \Phi''(\,.\,, \widetilde{x})$ in $(\mathcal{G}(\mathbb{R}))^n$. Consequently,

$$\Phi'(\eta, \widetilde{x}) = \Phi''(\eta, \widetilde{x})$$

$\forall\eta \in \mathcal{R}_c$ and all $\widetilde{x} \in \mathcal{R}^n$. The claim now follows from [27, Theorem 2.10]. □

As in the classical theory, we are first going to investigate symmetry groups of algebraic equations.

Definition 4.6. *Let $F \in \mathcal{G}_\tau(\mathbb{R}^n)$ and let Φ be a generalized group action on $\mathbb{R}^n$. Φ is called a symmetry group of the equation*

$$F(x) = 0$$

in $\mathcal{G}_\tau(\mathbb{R}^n)$ if for any $\widetilde{x} \in \mathcal{R}^n$ with $F(\widetilde{x}) = 0 \in \mathcal{R}$ it follows that $\eta \to F(\Phi(\eta, \widetilde{x})) = 0$ in $\mathcal{G}(\mathbb{R})$ (or, equivalently, $F(\Phi(\eta, \widetilde{x})) = 0$ in $\mathcal{R}$ for every $\eta \in \mathcal{R}_c$).

A characterization of symmetry groups of (generalized) algebraic equations in terms of infinitesimal generators is provided by the following theorem.

Theorem 4.7. *Let $F \in \mathcal{G}_\tau(\mathbb{R}^n)$ be of the form*

$$F(x_1, \dots, x_n) = x_i - f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

for some $1 \le i \le n$ and $f \in \mathcal{G}_\tau(\mathbb{R}^{n-1})$. Let Φ be a $\mathcal{G}$-complete generalized group action with infinitesimal generator $X = \sum_{i=1}^n \xi_i(x)\partial_{x_i}$ and suppose that $x' \to \xi(x', f(x'))$ defines a generalized vector field on $\mathbb{R}^{n-1}$ such that the corresponding system of ODEs possesses a flow in $(\widetilde{\mathcal{G}}_\tau(\mathbb{R}^{1+(n-1)}))^{n-1}$. The following conditions are equivalent:

(i) *Φ is a symmetry group of $F(x) = 0$.*

(ii) *If $\widetilde{x} \in \mathcal{R}^n$ with $F(\widetilde{x}) = 0 \in \mathcal{R}$, it follows that $X(F)(\widetilde{x}) = 0$ in $\mathcal{R}$.*

Proof. (i) ⇒ (ii): Consider the function $(\eta, x) \to F(\Phi(\eta, x)) \in \widetilde{\mathcal{G}}_\tau(\mathbb{R}^{1+n})$. We have

$$\frac{d}{d\eta} F(\Phi(\eta, x)) = \sum_{i=1}^n \frac{\partial F}{\partial x_i}(\Phi(\eta, x))\xi_i(\Phi(\eta, x)) = X(F)(\Phi(\eta, x)),$$

so that $\frac{d}{d\eta}\Big|_0 F(\Phi(\eta, x)) = X(F)(x)$ in $\mathcal{G}_\tau(\mathbb{R}^n)$. Let $\widetilde{x} \in \mathcal{R}^n$ such that $F(\widetilde{x}) = 0$. Then $F(\Phi(\,.\,, \widetilde{x})) = 0$ in $\mathcal{G}(\mathbb{R})$. Thus $\frac{d}{d\eta}\Big|_0 F(\Phi(\eta, \widetilde{x})) = 0$ in $\mathcal{R}$ which means that $X(F)(\widetilde{x}) = 0$ in $\mathcal{R}$.

(ii) ⇒ (i): We assume $F(x_1, \dots, x_n) = x_n - f(x_1, \dots, x_{n-1})$ and abbreviate $(x_1, \dots, x_{n-1})$ by x'. Our first claim is that

$$\xi_n(x', f(x')) = \sum_{j=1}^{n-1} \xi_j(x', f(x'))\partial_j f(x') \text{ in } \mathcal{G}_\tau(\mathbb{R}^{n-1}).$$

Indeed, if $\widetilde{x}' \in \mathcal{R}^{n-1}$ then $F(\widetilde{x}', f(\widetilde{x}')) = 0$ in $\mathcal{R}$. Hence $X(F)(\widetilde{x}', f(\widetilde{x}')) = 0$ in $\mathcal{R}$ for all $\widetilde{x}'$ by our assumption. Our claim now follows from [27, Theorem 2.7]. Consider the following system of ODEs in $\mathcal{G}_\tau$:

$$\begin{aligned} &\dot{x}_j(t) = \xi_j(x', f(x')), \quad (j = 1, \dots, n-1), \\ &x'(0) = \widetilde{a}' \in \mathcal{R}^{n-1}. \end{aligned}$$

By our assumption, this system has a flow $(\eta, a') \to (h_1(\eta, a'), \dots, h_{n-1}(\eta, a'))$ in $(\widetilde{\mathcal{G}}_\tau(\mathbb{R}^{1+(n-1)}))^{n-1}$. Set $g_n(\eta, a) := f(h_1(\eta, a'),\dots,h_{n-1}(\eta, a'))$. Then $g_n(0, a) = f(a')$ and

$$g(\eta, a) = (g_1(\eta, a), \dots, g_n(\eta, a)) := (h_1(\eta, a'), \dots, h_{n-1}(\eta, a'), g_n(\eta, a))$$

is in $(\widetilde{\mathcal{G}}_\tau(\mathbb{R}^{1+n}))^n$. If $\widetilde{a} \in \mathcal{R}^n$ then $F(g(\eta, \widetilde{a})) = 0$ in $\mathcal{R}$ $\forall \eta \in \mathcal{R}_c$. Therefore, if we can show that $g(\,.\,, \widetilde{a}) = \Phi(\,.\,, \widetilde{a})$ in $(\mathcal{G}(\mathbb{R}))^n$ $\forall \widetilde{a}$ with $F(\widetilde{a}) = 0$, the proof is completed. Now we have $\dot{g}_j(\eta, a) = \xi_j(g_1(\eta, a), \ldots, g_n(\eta, a))$ for $1 \leq j \leq n-1$ and

$$\begin{aligned} \dot{g}_n(\eta, a) &= \sum_{i=1}^{n-1} \frac{\partial f}{\partial x_i}(g_1(\eta, a), \ldots, g_{n-1}(\eta, a))\dot{g}_i(\eta, a) \\ &= \xi_n(g_1(\eta, a), \ldots, f(g_1(\eta, a), \ldots, g_{n-1}(\eta, a))) = \xi_n(g(\eta, a)). \end{aligned}$$

If $F(\widetilde{a}) = 0$ in $\mathcal{R}$, then $\widetilde{a}_n = f(\widetilde{a}')$, so that $g(0, \widetilde{a}) = (\widetilde{a}', f(\widetilde{a}')) = \widetilde{a} = \Phi(0, \widetilde{a})$. Thus $g(\,.\,, \widetilde{a})$ and $\Phi(\,.\,, \widetilde{a})$ solve the same initial value problem. Since X is $\mathcal{G}$-complete, the claim follows. □

4.2. Symmetries of differential equations. In this section we are going to apply the above results to symmetry groups of differential equations involving generalized functions. To this end, we will first have to define generalized group actions on generalized functions. Once we have done this, by a symmetry group of a differential equation we will again mean a group action that transforms solutions into other solutions. Thus, from now on we will exclusively consider group actions on some space $\mathbb{R}^p \times \mathbb{R}^q$ of independent and dependent variables.

Definition 4.8. *A generalized group action* $\Phi \in (\widetilde{\mathcal{G}}_\tau(\mathbb{R} \times \mathbb{R}^{p+q}))^{p+q}$ *is called projectable if it is of the form*

$$\Phi(\eta, (x, u)) = (\Xi_\eta(x), \Psi_\eta(x, u)),$$

where $\Xi \in (\widetilde{\mathcal{G}}_\tau(\mathbb{R} \times \mathbb{R}^p))^p$ *and* $\Psi \in (\widetilde{\mathcal{G}}_\tau(\mathbb{R} \times \mathbb{R}^{p+q}))^q$.

The group properties in this case read

$$\Xi_{\eta_1+\eta_2} = \Xi_{\eta_1} \circ \Xi_{\eta_2} \quad \text{in } \mathcal{G}_\tau(\mathbb{R}^p)\ \forall \eta_1, \eta_2 \in \mathcal{R}_c, \tag{4.1}$$

$$\Psi_{\eta_1+\eta_2}(x, u) = \Psi_{\eta_1}(\Xi_{\eta_2}(x), \Psi_{\eta_2}(x, u)) \quad \text{in } \mathcal{G}_\tau(\mathbb{R}^{p+q})\ \forall \eta_1, \eta_2 \in \mathcal{R}_c. \tag{4.2}$$

In particular, we have

$$\Xi_\eta \circ \Xi_{-\eta} = \text{id} \quad \text{in } \mathcal{G}_\tau(\mathbb{R}^p)\ \forall \eta \in \mathcal{R}_c. \tag{4.3}$$

An adaptation of Lie group analysis to spaces of distributions faces the following fundamental problem. The methods of classical Lie group analysis of differential equations are *geometric* in the sense that group action on functions is defined via graphs, but in classical distribution theory there is no means of defining graphs of distributions. However, due to the pointvalue characterization obtained in [27] this problem can be dealt with in a satisfactory manner within Colombeau algebras.

Definition 4.9. *Let* $U \in (\mathcal{G}(\mathbb{R}^p))^q$ *and* $V \in (\mathcal{G}_\tau(\mathbb{R}^p))^q$. *The graphs of* U *and* V *are defined as*

$$\begin{aligned} \Gamma_U &:= \{(\widetilde{x}, U(\widetilde{x})) : \widetilde{x} \in \mathcal{R}_c^p\}, \\ \Gamma_V &:= \{(\widetilde{x}, V(\widetilde{x})) : \widetilde{x} \in \mathcal{R}^p\}. \end{aligned}$$

It follows directly from [27, Theorems 2.4 and 2.7] that any generalized function is uniquely determined by its graph. Our next aim is to define generalized group actions on generalized functions. As in the classical case this is done geometrically, i.e., by transformation of graphs. The following result is immediate from the definitions.

Proposition 4.10. *Let $U \in (\mathcal{G}_\tau(\mathbb{R}^p))^q$ and let Φ be a projectable generalized group action on $\mathbb{R}^p \times \mathbb{R}^q$. Then $\Phi_\eta(\Gamma_U) = \Gamma_{\Phi_\eta(U)}$ in $\mathcal{R}^{p+q}$ for each η, where $\Phi_\eta(U)$ denotes the element*

$$x \to \Psi_\eta(\Xi_{-\eta}(x), U \circ \Xi_{-\eta}(x))$$

of $(\mathcal{G}_\tau(\mathbb{R}^p))^q$. □

We are now able to give a geometric characterization of solutions of PDEs in $\mathcal{G}_\tau$.

Proposition 4.11. *Consider the system of PDEs*

$$\Delta_\nu(x, U^{(n)}) = 0, \quad 1 \le \nu \le l, \tag{4.4}$$

in $\mathcal{G}_\tau(\mathbb{R}^p))^q$ (where $\Delta \in (\mathcal{G}_\tau((\mathbb{R}^p \times \mathbb{R}^q)^{(n)}))^l$). Set

$$\mathcal{S}_\Delta := \{\widetilde{z} \in \mathcal{R}^{(n)} \,:\, \Delta_\nu(\widetilde{z}) = 0 \ (1 \le \nu \le l)\}.$$

Then $U \in (\mathcal{G}_\tau(\mathbb{R}^p))^q$ is a solution of the system if and only if $\Gamma_{\mathrm{pr}^{(n)}U} \subseteq \mathcal{S}_\Delta$.

Proof. This follows immediately from [27, Theorem 2.7]. □

Prolongation of generalized group actions can be handled in a similar fashion as in the classical theory. Thus, let Φ be a projectable generalized group action on $\mathbb{R}^p \times \mathbb{R}^q$. We want to define the nth prolongation $\mathrm{pr}^{(n)}\Phi$ as a projectable generalized group action on $(\mathbb{R}^p \times \mathbb{R}^q)^{(n)}$. Let $z \in (\mathbb{R}^p \times \mathbb{R}^q)^{(n)}$ and choose $h \in \mathcal{O}_M(\mathbb{R}^p)^q$ such that $(z_1, \dots, z_p, \mathrm{pr}^{(n)}h(z_1, \dots, z_p)) = z$. Now set

$$\mathrm{pr}^{(n)}\Phi(\eta, z) := (\Xi_\eta(z_1, \dots, z_p), \mathrm{pr}^{(n)}(\Phi_\eta(h))(\Xi_\eta(z_1, \dots, z_p))). \tag{4.5}$$

Using for h a suitable Taylor polynomial, it follows that $\mathrm{pr}^{(n)}\Phi \in (\widetilde{\mathcal{G}}_\tau(\mathbb{R} \times (\mathbb{R}^p \times \mathbb{R}^q)^{(n)})^N$ (where $N = \dim((\mathbb{R}^{p+q})^{(n)})$). Moreover, the definition does not depend on the particular choice of h, which follows exactly as in the classical case.

Lemma 4.12. *Let $\widetilde{z} \in (\mathcal{R}^p \times \mathcal{R}^q)^{(n)}$ and assume that $U \in (\mathcal{G}_\tau(\mathbb{R}^p))^q$ satisfies $(\widetilde{z}_1, \dots, \widetilde{z}_p, \mathrm{pr}^{(n)}U(\widetilde{z}_1, \dots, \widetilde{z}_p)) = \widetilde{z}$. Then*

$$\mathrm{pr}^{(n)}\Phi(\eta, \widetilde{z}) = (\Xi_\eta(\widetilde{z}_1, \dots, \widetilde{z}_p), \mathrm{pr}^{(n)}(\Phi_\eta(U))(\Xi_\eta(\widetilde{z}_1, \dots, \widetilde{z}_p))) \quad \forall \eta \in \mathcal{R}_c. \tag{4.6}$$

Proof. Let $U = \mathrm{cl}[(u_\varepsilon)_{\varepsilon \in I}]$ and choose a representative $(z_\varepsilon)_{\varepsilon \in I}$ of $\widetilde{z}$ such that

$$(z_{1\varepsilon}, \dots, z_{p\varepsilon}, \mathrm{pr}^{(n)}u_\varepsilon(z_{1\varepsilon}, \dots, z_{p\varepsilon})) = z_\varepsilon \quad \forall \varepsilon.$$

Using the chain rule as in Proposition 3.7, it follows that the right hand sides of (4.5) (with z replaced by $\widetilde{z}$) and of (4.6) have the same representative (depending *exclusively* on $(z_\varepsilon)_{\varepsilon \in I}$). □

Proposition 4.13. *$\mathrm{pr}^{(n)}\Phi$ is a generalized group action on $(\mathbb{R}^p \times \mathbb{R}^q)^{(n)}$.*

Proof. Property 4.2 (i) is clearly satisfied. Concerning (ii), according to [27, Theorem 2.7] it suffices to show that

$$\mathrm{pr}^{(n)}\Phi(\eta_1 + \eta_2, \widetilde{z}) = \mathrm{pr}^{(n)}\Phi(\eta_1, \mathrm{pr}^{(n)}\Phi(\eta_2, \widetilde{z})) \quad \forall \eta_1, \eta_2 \in \mathcal{R}_c, \ \forall\, \widetilde{z} \in (\mathcal{R}^p \times \mathcal{R}^q)^{(n)}.$$

Choose some $U \in (\mathcal{G}_\tau(\mathbb{R}^p))^q$ with $(\widetilde{z}_1, \dots, \widetilde{z}_p, \mathrm{pr}^{(n)}U(\widetilde{z}_1, \dots, \widetilde{z}_p)) = \widetilde{z}$. Then due to Lemma 4.12 we have

$$\mathrm{pr}^{(n)}\Phi(\eta_2, \widetilde{z}) = (\Xi_{\eta_2}(\widetilde{z}_1, \dots, \widetilde{z}_p), \mathrm{pr}^{(n)}(\Phi_{\eta_2}(U))(\Xi_{\eta_2}(\widetilde{z}_1, \dots, \widetilde{z}_p))).$$

By (4.6) this implies $\mathrm{pr}^{(n)}\Phi(\eta_1, \mathrm{pr}^{(n)}\Phi(\eta_2, \widetilde{z})) = \mathrm{pr}^{(n)}\Phi(\eta_1 + \eta_2, \widetilde{z})$. □

As in the classical case we therefore have (using the notations from Proposition 4.11) the following proposition.

Proposition 4.14. *Let Φ be a projectable generalized group action on $\mathbb{R}^p \times \mathbb{R}^q$ such that $\mathrm{pr}^{(n)}\Phi$ is a symmetry group of the algebraic equation $\Delta(z) = 0$. Then Φ is a symmetry group of* (4.4).

Proof. If $U \in \mathcal{G}_\tau(\mathbb{R}^p)$ is a solution of (4.4), then $\Gamma_{\mathrm{pr}^{(n)}U} \subseteq \mathcal{S}_\Delta$ by Proposition 4.11. Thus

$$\Gamma_{\mathrm{pr}^{(n)}(\Phi_\eta U)} = \mathrm{pr}^{(n)}\Phi_\eta(\Gamma_{\mathrm{pr}^{(n)}U}) \subseteq \mathcal{S}_\Delta,$$

so that, again from Proposition 4.11, the claim follows. □

Definition 4.15. *Let X be a $\mathcal{G}$-complete generalized vector field. The nth prolongation of X is defined as the infinitesimal generator of the nth prolongation of the generalized group action Φ corresponding to X:*

$$\mathrm{pr}^{(n)}X\Big|_z = \frac{d}{d\eta}\Big|_0 \mathrm{pr}^{(n)}\Phi_\eta(z),$$

provided that $\mathrm{pr}^{(n)}\Phi$ is $\mathcal{G}$-complete as well. In this case, both X and Φ are called $\mathcal{G}$-n-complete.

From Theorem 4.7 and Proposition 4.14 we immediately conclude the following theorem.

Theorem 4.16. *Under the assumptions of Proposition* 4.11, *let Φ be a $\mathcal{G}$-n-complete generalized group action on $\mathbb{R}^p \times \mathbb{R}^q$ with infinitesimal generator X such that the conditions of Theorem* 4.7 *are satisfied for Δ and $\mathrm{pr}^{(n)}\Phi$. If*

$$\mathrm{pr}^{(n)}X(\Delta)(\widetilde{z}) = 0 \quad \forall \widetilde{z} \in (\mathcal{R}^p \times \mathcal{R}^q)^{(n)} \text{ with } \Delta(\widetilde{z}) = 0,$$

then Φ is a symmetry group of (4.4). □

In order to be able to apply the same algorithm as in classical Lie theory for the determination of the symmetry group of a generalized PDE, the final step is to verify that the formulas for prolongation of vector fields carry over to generalized vector fields.

Theorem 4.17. *Let*

$$X = (x, u) \to \sum_{i=1}^{p} \xi_i(x)\partial_{x_i} + \sum_{\alpha=1}^{q} \psi_\alpha(x, u)\partial_{u^\alpha}$$

be a $\mathcal{G}$-n-complete generalized vector field with corresponding projectable group action Φ on $(\mathbb{R}^p \times \mathbb{R}^q)$. Then

$$\mathrm{pr}^{(n)}X = X + \sum_{\alpha=1}^{q} \sum_J \psi_\alpha^J(x, u^{(n)})\partial_{u_J^\alpha},$$

where $J = (j_1, \dots, j_k)$, $1 \le j_k \le p$ for $1 \le k \le n$ and

$$\psi_\alpha^J(x, u^{(n)}) = D_J(\psi_\alpha - \sum_{i=1}^{p} \xi_i u_i^\alpha) + \sum_{i=1}^{p} \xi_i u_{J,i}^\alpha.$$

Proof. Using the machinery developed so far, this is an easy modification of the proof of the classical result (see [29, Theorem 2.36]). □

We may summarize the results of this section as follows: In order to determine the symmetries of a differential equation involving generalized functions, the algorithm (as in the classical case) is to make an ansatz for the infinitesimal generators, calculate their prolongations according to Theorem 4.17, and then use Theorem 4.16 to determine the defining equations for the coefficient functions of the infinitesimal generators. The defining equations now yield PDEs in $\mathcal{G}_\tau$. Any solution of these equations that defines a $\mathcal{G}$-n-complete generator will upon integration yield a symmetry group in $\mathcal{G}_\tau$.

Example 4.18. Scalar conservation laws of the form

$$u_t + F(u)u_x = 0 \tag{4.7}$$

arise in the kinetic theory of traffic flow. Here u denotes the density, and the propagation velocity F may be a strictly decreasing function of u with one or more jumps. A typical case is a unimodal flux function (whose derivative is F) with a kink at its maximum, as supported by experimental data [15]. Convolution with a nonnegative mollifier $(\rho_\varepsilon)_{\varepsilon\in I}$ allows us to interpret F as an element of $\mathcal{G}_\tau(\mathbb{R})$ which is invertible. Thus our theory of symmetry transformations for equations with generalized nonlinearities applies. The determining equations are

$$\begin{gathered}\varphi_t + F\varphi_x = 0,\\ -\xi_x + F\tau_t + \tau F_t + \varphi F_u - F\xi_x + F^2\tau_x + \xi F_x = 0\end{gathered}$$

with infinitesimal generator $\mathbf{v} = \xi(x,t)\partial_x + \tau(x,t)\partial_t + \varphi(x,t,u)\partial_u$. As a particular solution we obtain $\mathbf{v} = xt\partial_x + t^2\partial_t + (F'(u))^{-1}(x - tF(u))\partial_u$. The corresponding generalized group action can be calculated explicitly in $\mathcal{G}_\tau$ showing that if u is a $\mathcal{G}_\tau$ -solution to (4.7) then so is

$$(x,t) \to F^{-1}\left(\eta x(1+\eta t)^{-1} + F(u(x(1+\eta t)^{-1}, t(1+\eta t)^{-1})(1+\eta t)^{-1}\right).$$

In particular, a constant state u is transformed into a generalized solution to (4.7) which, depending on the shape of F, will generally be associated with a piecewise smooth function.

Example 4.19. The nonlinear d'Alembert–Hamilton system

$$\begin{gathered}u_{tt} - u_{xx} - u_{yy} - u_{zz} = F(u),\\ u_t^2 - u_x^2 - u_y^2 - u_z^2 = G(u)\end{gathered} \tag{4.8}$$

arises in the study of relativistic field equations [7] and as a constraint in reducing the nonlinear wave equation to an ODE [12, 13]. One of its symmetries is generated by the vector field $\mathbf{v} = \varphi(u)\partial_u$ where the function φ has to satisfy

$$\begin{gathered}F\varphi_u - \varphi F_u + G\varphi_{uu} = 0,\\ 2G\varphi_u - \varphi G_u = 0\,.\end{gathered}$$

In particular, in the isotropic case $F \equiv G \equiv 0$ the function φ is arbitrary. In our theory it may be taken in $\mathcal{G}_\tau(\mathbb{R})$ subject to the $\mathcal{G}$-completeness conditions formulated above. As an example of the possible behavior of generalized transformations, consider the vector field $\mathbf{v} = \varphi(u)\partial_u$ where $\varphi \in \mathcal{G}_\tau(\mathbb{R})$ is the class of $(\varphi_\varepsilon)_{\varepsilon\in I}$ with $\varphi_\varepsilon(u) = \tanh(\frac{u}{\varepsilon})$.

Thus $\varphi(u)$ is associated with the jump function $-\mathrm{sgn}(u)$. Starting with a classical smooth solution $u = u(x,t) \in \mathcal{O}_C(\mathbb{R}^4)$ of the isotropic d'Alembert–Hamilton system ((4.8) with $F \equiv G \equiv 0$), the generalized symmetry transform generated by the vector field $\mathbf{v}$ turns $u(x,t)$ into the generalized solution $\tilde{U} \in \mathcal{G}_\tau(\mathbb{R}^4)$ with representative

$$\tilde{u}_\varepsilon(x,t) = \varepsilon \operatorname{Arsinh}\left(e^{\eta/\varepsilon} \sinh \frac{u(x,t)}{\varepsilon} \right) .$$

When $\eta > 0$, it is straightforward to check that $\tilde{U}$ is associated with the piecewise smooth function $v(x,t) = u(x,t) + \eta\,\mathrm{sgn}(u(x,t))$. The generalized symmetry this way transforms smooth solutions into discontinuous solutions.

Acknowledgments. We would like to thank M. Grosser, G. Hörmann, and P. J. Olver for several helpful discussions. A number of constructive suggestions of the two referees led to improvements in the paper.

REFERENCES

[1] J. Aragona and H. A. Biagioni, *Intrinsic definition of the Colombeau algebra of generalized functions*, Anal. Math., 17 (1991), pp. 75–132.

[2] Y. Y. Berest, *Construction of fundamental solutions for Huygen's equations as invariant solutions*, Soviet Math. Dokl., 43 (1991), pp. 496–499.

[3] Y. Y. Berest, *Weak invariants of local groups of transformations*, Differential Equations, 29 (1993), pp. 1561–1567.

[4] Y. Y. Berest, *Group analysis of linear differential equations in distributions and the construction of fundamental solutions*, Differential Equations, 29 (1993), pp. 1700–1711.

[5] Y. Y. Berest and N. H. Ibragimov, *Group theoretic determination of fundamental solutions*, Lie Groups Appl., 1 (1994), pp. 65–80.

[6] H. A. Biagioni, *A Nonlinear Theory of Generalized Functions*, Lecture Notes in Math. 1421, Springer, Berlin, 1990.

[7] G. Ciecura and A. Grundland, *A certain class of solutions to the nonlinear wave equation*, J. Math. Phys., 25 (1984), pp. 3460–3469.

[8] J. F. Colombeau, *New Generalized Functions and Multiplication of Distributions*, North-Holland, Amsterdam, 1984.

[9] J. F. Colombeau, *Elementary Introduction to New Generalized Functions*, North-Holland, Amsterdam, 1985.

[10] J. F. Colombeau, *Multiplication of Distributions. A Tool in Mathematics, Numerical Engineering and Theoretical Physics*, Lecture Notes in Math. 1532, Springer, Berlin, 1992.

[11] L. E. Fraenkel, *Formulae for high derivatives of composite functions*, Math. Proc. Cambridge Philos. Soc., 83 (1978), pp. 159–165.

[12] W. I. Fushchish and R. Z. Zhdanov, *On some new exact solutions of the nonlinear d'Alembert-Hamilton system*, Phys. Lett. A, 141 (1989), pp. 113–115.

[13] W. I. Fushchish, R. Z. Zhdanov, and I. A. Yegorchenko, *On the reduction of the nonlinear multi-dimensional wave equations and compatibility of the d'Alembert-Hamilton system*, J. Math. Anal. Appl., 161 (1991), pp. 352–360.

[14] M. Grosser, M. Kunzinger, R. Steinbauer, and J. A. Vickers, *A Global Theory of Algebras of Generalized Functions*, Erwin Schrödinger International Institute of Mathematical Physics, Preprint ESI 813, Vienna, 1999.

[15] F. L. Hall, B. L. Allen, and M. A. Gunter, *Empirical analysis of freeway flow-density relationships*, Transpn. Res. A, 20 (1986), p. 197.

[16] R. Hermann and M. Oberguggenberger, *Ordinary differential equations and generalized functions*, in Nonlinear Theory of Generalized Functions, M. Grosser, G. Hörmann, M. Kunzinger, and M. Oberguggenberger, eds., Chapman and Hall/CRC, Boca Raton, 1999, pp. 85–98.

[17] L. Hörmander, *The Analysis of Linear Partial Differential Operators* II, Grundlehren Math. Wiss. 257, Springer-Verlag, Berlin, 1990.

[18] N. H. Ibragimov, *Group theoretical treatment of fundamental solutions*, in Physics on Manifolds, M. Flato, R. Kerner, and A. Lichnerowicz, eds., Kluwer, Dordrecht, 1994, pp. 161–175.

[19] N. H. Ibragimov, ed., *CRC Handbook of Lie Group Analysis of Differential Equations*, Vol. 1–3, CRC Press, Boca Raton, 1994–1996.

[20] M. Kunzinger, *Lie-Transformation Groups in Colombeau Algebras*, Doctoral thesis, University of Vienna, 1996.

[21] M. Kunzinger and M. Oberguggenberger, *Symmetries of differential equations in Colombeau algebras*, in Modern Group Analysis VI, N. H. Ibragimov and F. M. Mahomed, eds., New Age Int. Publ., New Delhi, 1997, pp. 9–20.

[22] P. D. Methée, *Sur les distributions invariantes dans le groupe des rotations de Lorentz*, Comment. Math. Helv., 28 (1954), pp. 224–269.

[23] E. Michael, *Continuous selections* I, Ann. Math. (2), 63 (1956), pp. 361–382.

[24] M. Oberguggenberger, *Multiplication of Distributions and Applications to Partial Differential Equations*, Pitman Res. Notes Math. 259, Longman, Harlow, UK, 1992.

[25] M. Oberguggenberger, *Case study of a nonlinear, nonconservative, nonstrictly hyperbolic system*, Nonlinear Anal., 19 (1992), pp. 53–79.

[26] M. Oberguggenberger, *Nonlinear theories of generalized functions*, in Advances in Analysis, Probability, and Mathematical Physics-Contributions from Nonstandard Analysis, S. Albeverio, W. A. J. Luxemburg, and M. P. H. Wolff, eds., Kluwer, Dordrecht, 1994, pp. 56–74.

[27] M. Oberguggenberger and M. Kunzinger, *Characterization of Colombeau generalized functions by their pointvalues*, Math. Nachr., 203 (1999), pp. 147–157.

[28] M. Oberguggenberger and E. E. Rosinger, *Solution of Continuous Nonlinear PDEs Through Order Completion*, North-Holland, Amsterdam, 1994.

[29] P. J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed., Springer, New York, 1993.

[30] E. E. Rosinger and M. Rudolph, *Group invariance of global generalized solutions of smooth nonlinear PDEs: A Dedekind order completion method*, Lie Groups Appl., 1 (1994), pp. 203–215.

[31] E. E. Rosinger and Y. E. Walus, *Group invariance of generalized solutions obtained through the algebraic method*, Nonlinearity, 7 (1994), pp. 837–859.

[32] L. Schwartz, *Sur l'impossibilite de la multiplication des distributions*, C.R. Acad. Sci. Paris, 239 (1954), pp. 847–848.

[33] Z. Szmydt, *On homogeneous rotation invariant distributions and the Laplace operator*, Ann. Polon. Math., 6 (1979), pp. 249–259.

[34] Z. Szmydt, *Fourier Transformation and Linear Partial Differential Equations*, D. Reidel, Dordrecht, 1977.

[35] Z. Szmydt and B. Ziemian, *Invariant fundamental solutions of the wave operator*, Demonstratio Math., 19 (1986) pp. 371–386.

[36] A. Tengstrand, *Distributions invariant under an orthogonal group of arbitrary signature*, Math. Scand., 8 (1960), pp. 201–218.

[37] J.A. Vickers, *Nonlinear generalized functions in general relativity*, in Nonlinear Theory of Generalized Functions, M. Grosser, G. Hörmann, M. Kunzinger, and M. Oberguggenberger, eds., Chapman and Hall/CRC, Boca Raton, 1999, pp. 275–290.

[38] B. Ziemian, *On distributions invariant with respect to some linear transformations*, Ann. Polon. Math., 6 (1979), pp. 261–276.

SIAM J. Math. Anal.
Vol. 31, No. 6, pp. 1214–1234

DISCRETE AND CONTINUOUS DIRICHLET-TO-NEUMANN MAPS IN THE LAYERED CASE*

DAVID V. INGERMAN†

Abstract. Every sufficiently regular nonnegative function γ (conductivity) on the closed unit disk $\overline{\mathbb{D}}$ induces the *Dirichlet-to-Neumann map* Λ_γ on functions on $\partial\mathbb{D}$. The main inverse problems are to give a characterization of the maps Λ_γ and to find out when Λ_γ uniquely determines γ. In this paper we consider the case of conductivities that are constant on circles centered at the origin, and a discrete analogue of this so called *layered case*. We characterize a closure of the set of the layered Dirichlet-to-Neumann maps in terms of their kernels and spectra. We give sharp conditions for the uniqueness in the discrete inverse problem, and conditions on γ for the uniqueness in the continuous problem that we conjecture to be sharp. The characterization in terms of the spectra shows that continuous Dirichlet-to-Neumann maps can be viewed as limits of the discrete Dirichlet-to-Neumann maps. The characterization in terms of the kernels supports the conjecture in [D. Ingerman and J. Morrow, *SIAM J. Math. Anal.*, 29 (1998), pp. 106–115] that the *alternating property* essentially characterizes continuous Dirichlet-to-Neumann maps. The characterizations above give a new interpretation of connections between positive measures, positive definite functions, and analytic functions that map the right half-plane to itself in the Bochner and Herglotz theorems.

Key words. inverse problems, Dirichlet-to-Neumann map, Pick–Nevalinna interpolation, vibrations of inhomogeneous string

AMS subject classification. 35R30

PII. S0036141097326581

1. Introduction.

We first give a short outline of the paper. Precise definitions and theorems are in the next section.

Every sufficiently smooth positive function γ on the closed unit disk $\overline{\mathbb{D}}$ induces the Dirichlet-to-Neumann map Λ_γ from functions on the boundary of the disk $\partial\mathbb{D}$ to functions on $\partial\mathbb{D}$. There are two main inverse problems connected with the maps Λ_γ:

- The characterization of the Dirichlet-to-Neumann maps. (A necessary condition is known: it was shown in [8] and [4] that if $\gamma \in C^2(\overline{\mathbb{D}})$ then Λ_γ has so called alternating property, which can be thought of as a generalized Hopf's lemma property.)
- The problem of finding γ from Λ_γ. (A sufficient condition for uniqueness is known: It was recently proved in [13] that Λ_γ uniquely determines γ for $\gamma \in W^{2,p}(\overline{\mathbb{D}}) \subset C(\overline{\mathbb{D}})$, $p > 1$.)

Great progress in understanding of the discrete analogues (Dirichlet-to-Neumann maps on graphs) of these problems was made in [3] and [2]: it was shown that the discrete analogue of the alternating property essentially characterizes the discrete Dirichlet-to-Neumann maps. Also the discrete inverse problem was completely solved. The success in understanding the discrete problems gives one of the main motivations of this paper: to show a strong connection between properties of the continuous and discrete Dirichlet-to-Neumann maps.

In this paper we consider the case of conductivities that are constant on circles centered at the origin. We also consider a discrete analogue of this so called layered

*Received by the editors August 29, 1997; accepted for publication (in revised form) February 24, 1999; published electronically May 11, 2000. This work was supported by an Alfred P. Sloan Research Fellowship.

http://www.siam.org/journals/sima/31-6/32658.html

†Department of Mathematics, University of Washington, Seattle, WA 98195 (ingerman@math.washington.edu).

case. We obtain a clear picture of the sets of both discrete and continuous Dirichlet-to-Neumann maps in this case. We characterize their closure in terms of their kernels and spectra.

The characterization in terms of the spectra shows that continuous Dirichlet-to-Neumann maps can be viewed as limits of the discrete Dirichlet-to-Neumann maps.

The characterization in terms of the kernels supports the conjecture in [8] that the alternating property essentially characterizes continuous Dirichlet-to-Neumann maps.

We also conjecture sharp conditions on γ for the uniqueness in the continuous inverse problem. For the discrete case, we give a new algorithm, based on the Pick–Nevalinna interpolation theorem, for the recovery of γ.

The characterizations above give a new interpretation of the connection between positive measures, positive definite functions, and analytic functions from the right half-plane to itself: these objects describe, respectively, spectral measures, kernels, and spectra of Dirichlet-to-Neumann maps in the layered case.

2. Background and main results.

2.1. Basic definitions. We first give the definition of Dirichlet-to-Neumann maps as it is usually done; see [17] for details. For the layered case, which we will consider, the restrictions on γ will be weakened, and the domain of Λ_γ will be shrunk.

Let $\gamma \in C^{1,1}(\overline{\mathbb{D}})$. A function $u \in H^1(\mathbb{D})$ is called a γ-*harmonic* function or *potential* if

$$\operatorname{div}(\gamma \nabla u) = 0 \text{ in } \mathbb{D}. \tag{2.1}$$

A potential in $\mathbb{D}$ satisfies this equation if there are no sources or sinks of current in $\mathbb{D}$.

For each $f \in H^{1/2}(\partial\mathbb{D})$ there exists a unique γ-harmonic function u such that $u|_{\partial\mathbb{D}} = f$. The Dirichlet-to-Neumann corresponding to γ maps the boundary values of a γ-harmonic function (Dirichlet data) to the current flux $\gamma\frac{\partial u}{\partial r}|_{r=1}$ at the boundary (Neumann data). In symbols

$$\Lambda_\gamma = \gamma\frac{\partial u}{\partial r}|_{r=1},$$

where u is γ-harmonic and $u|_{\partial\mathbb{D}} = f$. The operator $\Lambda_\gamma : H^{1/2}(\partial\mathbb{D}) \to H^{-1/2}(\partial\mathbb{D})$ is a self-adjoint pseudodifferential operator of order 1.

A discrete analogue of the disk $\mathbb{D}$ is a *circular planar* graph. It is a finite graph $\Gamma = (V, E, \partial\Gamma)$ imbedded into $\overline{\mathbb{D}}$, where the set V is the set of nodes of the graph, the set E is the set of edges of Γ, and $\partial\Gamma = V \cap \partial\mathbb{D}$ is the nonempty set of *boundary* nodes of Γ. The set $V - \partial\Gamma$ is the set of *interior* nodes of Γ. A conductivity γ is a positive function on the edges of Γ.

A function u on the nodes of Γ is γ-harmonic if at every interior node p it satisfies Kirchhoff's law, that is, *the total current* $I_u(p)$ *out of* p *is zero*:

$$I_u(p) = \sum_{pq \in E} \gamma(pq)(u(p) - u(q)) = 0. \tag{2.2}$$

This tells that the value of u at p is the weighted average of the values of u at the neighbors of p (neighbors are the nodes q of the graph for which $pq \in E$). It follows that γ-harmonic functions satisfy the maximum and the minimum principles. From now on we will consider only the graphs in which every interior node is topologically connected to at least one boundary node. On such graphs (and only on them) each

γ-harmonic function u is uniquely determined by its values $u|_{\partial\Gamma}$ on the boundary of Γ. The discrete Dirichlet-to-Neumann map Λ_γ is the linear map that sends the boundary values f of a γ-harmonic function u to the corresponding total current out of nodes at the boundary $I_u|_{\partial\Gamma}$. Or algebraically,

$$\Lambda_\gamma f(p) = I_u(p) = \sum_{pq\in E} \gamma(pq)(u(p) - u(q)), p \in \partial\Gamma, \tag{2.3}$$

where u is γ-harmonic and $u|_{\partial\Gamma} = f$.

2.2. Alternating property. One of the main motivations of this paper is to show a strong connection between properties of discrete and continuous Dirichlet-to-Neumann maps. An important step in this direction has been made in [8], [3], and [4], where it was shown that both discrete and continuous Dirichlet-to-Neumann maps have the *alternating property.*

THEOREM 2.1 (see [8] and [4]). *Let Λ_γ be a Dirichlet-to-Neumann map for $\gamma \in C^2(\overline{\mathbb{D}})$. Then Λ_γ has the alternating property. That is, let A and B be a pair of disjoint intervals on $\partial\mathbb{D}$ and $f \in C^\infty(\partial\mathbb{D})$, such that* supp$f \subset A$. *Then for any m distinct points $b_1, b_2, \dots, b_m \in B$, numbered* clockwise*, such that*

$$(-1)^i \Lambda f(b_i) > 0,$$

there exists m distinct points $a_1, a_2, \dots, a_m \in A$ numbered counterclockwise*, such that*

$$(-1)^i f(a_i) < 0.$$

Figure 2.1 shows the main idea of the proof: the pattern of the directions of the current flux $\Lambda_\gamma f$ on B together with the maximum and minimum principles guarantee the existence of nonintersecting curves from b's on which potential alternates in sign. Since supp$f \subset A$ these curves have to terminate at A. See [8] for a detailed proof.

The same argument can be applied to the discrete case. In fact the discrete version of the alternating property, which we define next, essentially characterizes the discrete Dirichlet-to-Neumann maps.

We identify the space of real functions on $\partial\Gamma$ with $\mathbb{R}^n$, where n is the number of points in $\partial\Gamma$.

THEOREM 2.2 ([3]). *A self-adjoint linear map $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ is a Dirichlet-to-Neumann map of a circular planar graph if and only if $\Lambda 1 = 0$ and Λ has the discrete alternating property, that is, let A, B be a pair of disjoint intervals on $\partial\mathbb{D}$ and f a function on $\partial\Gamma$ with* supp$f \subset A$. *Then for any m points ($2m \le n$) $b_1, b_2, \dots, b_m \in B \cap \partial\Gamma$, numbered clockwise and such that*

$$(-1)^i \Lambda f(b_i) > 0,$$

there exist m distinct points $a_1, a_2, \dots, a_m \in A \cap \partial\Gamma$, numbered counterclockwise such that

$$(-1)^i f(a_i) < 0.$$

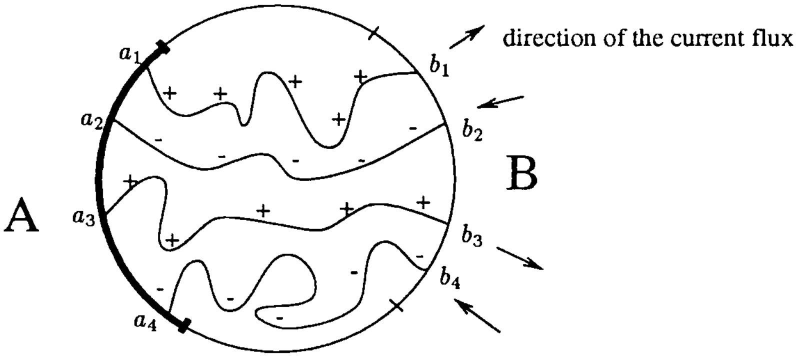


FIG. 2.1.

2.3. Right sign property. The following algebraic description of the alternating property turned out to be very useful. To state it we consider the *kernel* of Λ_γ.

For a discrete Dirichlet-to-Neumann map Λ_γ its kernel is the matrix that represents the linear operator Λ_γ.

The kernel of a continuous Dirichlet-to-Neumann map is the distribution $K(\phi, \theta)$ on $\partial\mathbb{D} \times \partial\mathbb{D}$ such that

$$\Lambda_\gamma f(\phi) = \int_0^{2\pi} K(\phi, \theta) f(\theta) d\theta. \tag{2.4}$$

The existence of K is guaranteed by the fact that Λ_γ is a pseudodifferential operator. In fact for $\gamma \in C^2(\overline{\mathbb{D}})$ K is a continuous function off the diagonal of $\partial\mathbb{D} \times \partial\mathbb{D}$, and the singularity at the diagonal is of order 2 (see [8]). The following theorem shows the equivalence of the alternating property of an operator Λ and the algebraic property of the kernel of Λ.

THEOREM 2.3 ([3]). *A symmetric matrix Λ is the kernel of a linear operator $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ that has the alternating property if and only if the kernel of Λ has the discrete* right sign *property. That is, for any two disjoint intervals $A, B \subset \partial\mathbb{D}$ and any $2m$ ($2m \leq n$) distinct points $a_1, a_2, \ldots, a_m \in A \cap \partial\Gamma$, $b_1, b_2, \ldots, b_m \in B \cap \partial\Gamma$ (as before a's are numbered counterclockwise and b's are numbered clockwise),*

$$\det\{-\Lambda(b_i, a_j)\}_{i,j=1}^m \geq 0.$$

A continuous analogue of this theorem was recently proved in [8].

THEOREM 2.4. *Let K be a distribution on $\partial\mathbb{D} \times \partial\mathbb{D}$ such that K is continuous off the diagonal and has a singularity of order 2 on the diagonal. Then the operator*

$$\Lambda f = \int K f$$

has the alternating property if and only if K has the continuous right sign *property. That is, for any two disjoint intervals $A, B \subset \partial\mathbb{D}$ and any $2m$ distinct points $a_1, a_2, \ldots, a_m \in A$, $b_1, b_2, \ldots, b_m \in B$ (a's and b's are numbered as above),*

$$\det\{-K(a_i, b_j)\}_{i,j=1}^m > 0.$$

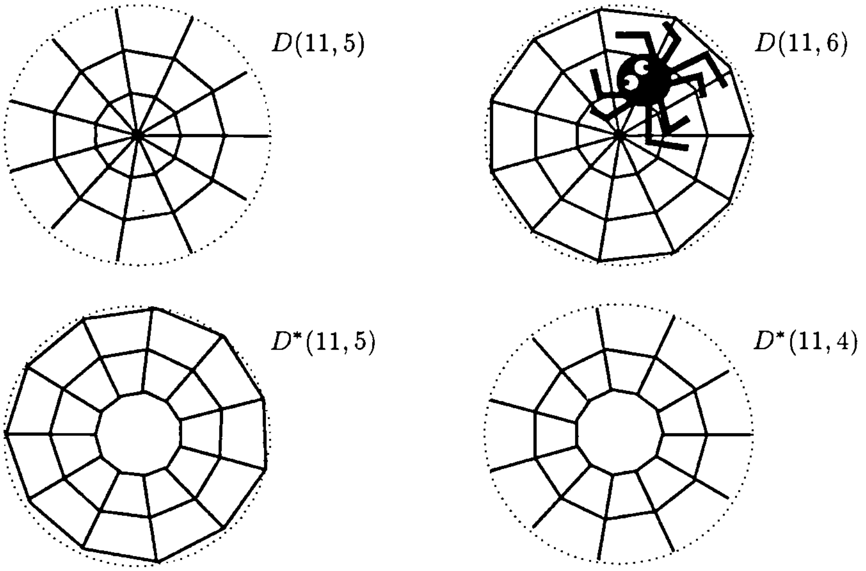


FIG. 2.2.

2.4. The layered case and the admittance function. In this paper we consider the case of conductivities that are constant on circles centered at the origin.

We now introduce a discrete analogue of the continuous layered situation. The *discrete disks* are connected circular planar graphs $D(n,l)$ and $D^*(n,l)$ of the following shapes shown in Figure 2.2, where n is the number of radial lines and l is the number of *layers*. The layers of the graphs $D(n,l)$ and $D^*(n,l)$ are minimal subsets of edges invariant under rotations of the graph by the angle $\frac{2\pi}{n}$. Each layer consists of n edges. We assume that the conductivity γ is constant on layers. Therefore, the layered conductivity is determined by l positive numbers.

We first describe the effect that the assumed form of γ has on Λ_γ. In both discrete and continuous situations the γ-harmonic functions are still γ-harmonic after rotations and reflections with respect to the origin. Therefore, the discrete and continuous Dirichlet-to-Neumann maps, corresponding to the layered conductivities γ, commute with rotations and the reflections of functions on the boundaries. For the continuous case it immediately follows that Dirichlet-to-Neumann maps commute with the Laplacian on the boundary of the disk $\frac{d^2}{d\theta^2}$. With a little more effort one gets that $\Lambda_\gamma 1 = 0$ and

$$\Lambda_\gamma e^{\pm ik\theta} = R(k)e^{\pm ik\theta}, k \in \mathbb{N}. \tag{2.5}$$

We call the function R the *admittance* function. Its values at positive integers uniquely determine Λ_γ. Let C be the set of positive measurable conductivities γ on $[0,1]$ such that for all $\epsilon > 0$

$$\int_\epsilon^1 \left(\gamma + \frac{1}{\gamma}\right) dr < \infty.$$

The results of Krein and Kac [9] (see also [5]) show that the admittance functions and therefore the Dirichlet-to-Neumann maps are well-defined on C. Let L be the set of Dirichlet-to-Neumann maps for conductivities from C. We will give it a weak topology defined by

$$\Lambda_n \to \Lambda \Leftrightarrow \Lambda_n e^{ik\theta} \to \Lambda e^{ik\theta} \Leftrightarrow R_n(|k|) \to R(|k|) \text{ for all } k \in \mathbb{Z}.$$

We will slightly abuse terminology, talking about admittance functions corresponding to the maps from $\overline{L}$.

We now will make sense of the admittance function for the discrete Dirichlet-to-Neumann maps. The discrete version of the Laplacian on the boundary of a discrete disk $D(n,l)$ or $D^*(n,l)$ is given by the $n \times n$ matrix of the form

$$\left[\frac{d^2}{d\theta^2}\right] = -\begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ -1 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix}.$$

It makes the calculations cleaner if we assume that n is odd. Throughout this section we let $n = 2m+1, m \in \mathbf{N}$. We define

$$\partial_n = \left\{\frac{2\pi j}{n}, j = -m, \dots, 0, \dots, m\right\}. \tag{2.6}$$

Direct calculation shows that $[\frac{d^2}{d\theta^2}]$ is diagonal in the orthogonal basis

$$e^{ik\theta}|_{\partial_n}, k = -m, \dots, 0, \dots, m$$

with the eigenvalues

$$-|e^{i\frac{2\pi k}{n}} - 1|^2, k = -m, \dots, 0, \dots, m.$$

We define

$$\omega_k^{(n)} = \omega_{-k}^{(n)} = |e^{i\frac{2\pi k}{n}} - 1|, k = -m, \dots, 0, \dots, m. \tag{2.7}$$

Note that

$$\lim_{n\to\infty} \frac{n}{2\pi}\omega_k^{(n)} = |k|.$$

The discrete Λ_γ commute with $[\frac{d^2}{d\theta^2}]$ and we get that the eigenvectors of the discrete Dirichlet-to-Neumann maps are the restrictions of the eigenfunctions of the continuous Dirichlet-to-Neumann maps to the boundaries of the discrete disks. In symbols,

$$\Lambda_\gamma e^{\pm ik\theta}|_{\partial_n} = R(\omega_k^{(n)}) e^{\pm ik\theta}|_{\partial_n}, k = 1, \dots, m. \tag{2.8}$$

2.5. Characterization of admittance functions. Now, to see how “close” the discrete and continuous Dirichlet-to-Neumann maps are, we need to describe their possible eigenvalues. We will give the descriptions by characterizing the discrete and continuous admittance functions.

The discrete admittance functions will turn out to be of the form of the Stieltjes' continued fractions:

$$R(\lambda) = \cfrac{1}{\cfrac{1}{\gamma_l} + \cfrac{1}{\gamma_{l-1}\lambda^2 + \cdots + \cfrac{1}{\cfrac{1}{\gamma_3} + \cfrac{1}{\gamma_2\lambda^2 + \cfrac{1}{\cfrac{1}{\gamma_1}}}}}}, \tag{2.9}$$

where γ_i's are the conductivities on the layers of the discrete disks. This explicit formula will allow us to show a one-to-one correspondence between the admittance functions of discrete disks with l layers and the Blaschke products of degree l. (This correspondence together with the Pick–Nevanlinna interpolation theorem is a key to the discrete inverse problem; see section 4.3.)

It follows from (2.9) that

$$\beta(\lambda) = \frac{R(\lambda)}{\lambda} \tag{2.10}$$

has a natural extension to an analytic function from the right half-plane $\mathbb{C}^+$ to itself. We define

$$\mathfrak{B} = \{\beta : \mathbb{C}^+ \to \mathbb{C}^+,\ \beta \text{ is analytic},\ \beta(\lambda) > 0 \text{ for } \lambda > 0\}.$$

Applying the Pick–Nevanlinna interpolation theorem (see [12], [16]), we will prove the following theorem.

THEOREM 2.5. *A linear map $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ is the Dirichlet-to-Neumann map of a discrete disk if and only if Λ is diagonal in the orthogonal basis*

$$e^{ik\theta}|_{\partial_n}, k = -m, \ldots, 0, \ldots, m,$$

$\Lambda 1 = 0$*, and there is a function β in $\mathfrak{B}$ such that*

$$\Lambda e^{\pm ik\theta}|_{\partial_n} = \omega_k^{(n)} \beta(\omega_k^{(n)}) e^{\pm ik\theta}|_{\partial_n}, k = 1, \ldots, m.$$

In other words the set of the Dirichlet-to-Neumann maps is equal to

$$\left\{ \sqrt{-\left[\frac{d^2}{d\theta^2}\right]} \beta \left(\sqrt{-\left[\frac{d^2}{d\theta^2}\right]} \right) : \beta \in \mathfrak{B} \right\}. \tag{2.11}$$

It turns out that a continuous analogue of this theorem is true. (Our proof heavily uses the characterization of spectral measures of inhomogeneous strings done by Krein and Kac [9]; see also [5].)

THEOREM 2.6. *The closure of the set of the layered Dirichlet-to-Neumann maps for conductivities in C equals*

$$\overline{L} = \left\{ \sqrt{-\frac{d^2}{d\theta^2}} \beta \left(\sqrt{-\frac{d^2}{d\theta^2}} \right) : \beta \in \mathfrak{B} \right\}. \tag{2.12}$$

Remark. Equation (2.1) in the layered case after separation of variables is sometimes replaced by the integral equation described in section 3.2 (see [14]). This allows one to define the admittance function, and therefore the Dirichlet-to-Neumann map, for conductivity given by a positive measure. This may be a cleaner setting in which one does not need to take the closure of the set of Dirichlet-to-Neumann maps to get the equality above.

2.6. The "$\gamma \leftrightarrow \frac{1}{\gamma}$" duality.

We note, without a proof, that the following identity is true; see [9] or [5]:

$$\Lambda_\gamma \Lambda_{\frac{1}{\gamma}} = \Lambda_{\frac{1}{\gamma}} \Lambda_\gamma = -\frac{d^2}{d\theta^2}.$$

DEFINITION 2.7. *Two disks $D_\gamma(n,l)$ and $D^*_{\frac{1}{\gamma}}(n,l)$ with conductivities on layers, respectively, $\{\delta_1, \xi_1, \delta_2, \xi_2, \delta_3, \dots\}$ and $\{\frac{1}{\delta_1}, \frac{1}{\xi_1}, \frac{1}{\delta_2}, \frac{1}{\xi_2}, \frac{1}{\delta_3}, \dots\}$ are called* dual.

We will show that

$$\Lambda(D^*_{\frac{1}{\gamma}})\Lambda(D_\gamma) = \Lambda(D_\gamma)\Lambda(D^*_{\frac{1}{\gamma}}) = -\left[\frac{d^2}{d\theta^2}\right].$$

Remark. We find the following result amusing. It was motivated by a question of G. Uhlmann. This question, together with the paper [16], has stimulated this paper. If γ is identically 1 on $\overline{\mathbb{D}}$ then the corresponding Dirichlet-to-Neumann map as an operator is the positive square root of the minus Laplacian on $\partial\mathbb{D}$. In symbols,

$$\Lambda_1 = \sqrt{-\frac{d^2}{d\theta^2}}.$$

The question: *Is*

$$\sqrt{-\left[\frac{d^2}{d\theta^2}\right]}$$

the Dirichlet-to-Neumann map of a circular planar graph? The answer is yes. It is an easy corollary of Theorem 2.5.

2.7. Approximation of continuous Dirichlet-to-Neumann maps by discrete ones.

Remark. We note that from the results in [8] it follows that the discrete disks give all possible Dirichlet-to-Neumann maps of circular planar graphs. In particular, the discrete disks with layered conductivity give all possible Dirichlet-to-Neumann maps Λ of circular planar graphs with the eigenvectors $e^{\pm ik\theta}|_{\partial_n}$. Therefore, for the purposes of the approximations of continuous Dirichlet-to-Neumann maps of the disk with layered conductivity by the discrete ones, one loses nothing essential considering only the discrete disks with layered conductivity and not all circular planar graphs.

The Pick–Nevalinna interpolation theorem (see [12]) lets us formulate the following "continuous is the limit of discrete" theorem.

THEOREM 2.8. *In the layered case, the eigenvectors of discrete Dirichlet-to-Neumann maps are restrictions of the eigenfunctions of the continuous Dirichlet-to-Neumann maps.*

Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the first (corresponding to $e^{i\theta}, e^{2i\theta}, \ldots, e^{ik\theta}$) eigenvalues of a map from $\overline{L}$. Then there exists a sequence of discrete disks $\{D_n\}$ with the first eigenvalues $\lambda_1^n, \lambda_2^n, \ldots, \lambda_k^n$ such that

$$\lambda_j = \lim_{n\to\infty} \lambda_j^n, 1 \leq j \leq k.$$

Conversely, let $\{D_n\}$ be a sequence of discrete disks such that the limits above exist. Then there exists an element of $\overline{L}$ with the first k eigenvalues being equal to the limits.

2.8. Characterization of kernels: Positive definite functions. We will now show the existence and characterize the kernels of the Dirichlet-to-Neumann maps in the layered case. We show the existence by an explicit calculation of the kernel of a map from $\overline{L}$ in terms of the corresponding admittance function $R(\lambda)$. Recall that the layered Dirichlet-to-Neumann maps commute with rotations and reflections (with respect to the origin) of functions. It follows that the kernel of Λ_γ has to be of the convolution type

$$K(\phi, \theta) = h(\phi - \theta),$$

where h is a distribution on $\mathbb{R}$ such that

$$h(s) = h(s + 2\pi) = h(2\pi - s), s \in \mathbb{R}$$

and

$$\int_0^{2\pi} h(s) \cos \lambda s ds = R(\lambda), \lambda > 0.$$

To proceed we need the following representation of analytic functions that maps the right half-plane to itself (see [1]).

Theorem 2.9 (Herglotz). *A function β is in $\mathfrak{B}$ if and only if for some $c, C \geq 0$*

$$\beta(\lambda) = C\lambda + \frac{c}{\lambda} + \int_0^\infty \frac{\lambda(1+t^2)d\sigma(t)}{\lambda^2 + t^2},$$

where σ is a positive measure of bounded variation on $(0, \infty)$.

A straightforward calculation gives us the following lemma.

Lemma 2.10. *A distribution K on $\partial\mathbb{D} \times \partial\mathbb{D}$ is the kernel of a map in the closure of the set of the layered Dirichlet-to-Neumann maps if and only if*

$$K(\phi, \theta) = h(\phi - \theta),$$

where h is a distribution on $\mathbb{R}$ such that $h(s) = h(s + 2\pi) = h(2\pi - s), s \in \mathbb{R}$,

$$\int_0^{2\pi} h(s) ds = 0,$$

and for $s \in [0, 2\pi)$

$$h(s) = c\delta(0) - C\delta''(0) + \int_0^\infty t\frac{e^{-st} + e^{(s-2\pi)t}}{1 - e^{-2\pi t}}(1+t^2)d\sigma(t), \tag{2.13}$$

where $c, C \geq 0$ and σ is a positive measure of bounded variation on $(0, \infty)$.

Corollary 2.11. *The kernel of a layered Dirichlet-to-Neumann map is C^∞ off the diagonal.*

We will now explain a connection of the characterization in Lemma 2.10 with the alternating property.

Definition 2.12. *A continuous function f on a possibly infinite interval (a, b) is* positive definite *if*

$$\det\{f(x_i + y_j)\}_1^m \geq 0$$

for all $m \in \mathbb{N}$, $x_i + y_j \in (a, b)$.

It follows that f is positive definite on $(0, 2\pi)$ if and only if the kernel $K(\phi, \theta) = f(\phi - \theta)$ satisfies the right sign property. We are now one step from restating the characterization of the kernels in terms of their right sign property. We need the following classical characterization of the positive definite functions (see [10]).

Theorem 2.13 (Bochner). *A continuous function f is positive definite on a possibly infinite interval (a, b) if and only if there exists a positive σ-finite measure ν on $\mathbb{R}$ such that*

$$f(x) = \int_{-\infty}^{+\infty} e^{xt} d\nu(t).$$

We now state one of the main results of this paper.

Theorem 2.14. *A distribution K on $\partial\mathbb{D} \times \partial\mathbb{D}$ is the kernel of a map in the closure of the set of the layered Dirichlet-to-Neumann maps if and only if*

$$K(\phi, \theta) = h(\phi - \theta),$$

where h is a distribution on $\mathbb{R}$ such that $h(s) = h(s + 2\pi) = h(2\pi - s), s \in \mathbb{R}$,

$$\int_0^{2\pi} h(s) ds = 0, \qquad \int_0^{2\pi} h(s)(\cos s - 1) ds < \infty,$$

and h is positive definite on $(0, 2\pi)$.

A discrete analogue of this theorem is an easy corollary of Theorems 2.2 and 2.3 and conductivity recovery algorithm in [3].

2.9. The inverse problems.

The continuous inverse problem will be reduced to an inverse Sturm–Liouville problem studied by Krein; see [9]. We obtain the following result.

Theorem 2.15. *The map from layered conductivities to corresponding Dirichlet-to-Neumann maps is injective on C.*

Conjecture. We would conjecture that the condition above for uniqueness in the inverse problem is sharp since if for some $\epsilon > 0$

$$\int_\epsilon^1 \left(\gamma + \frac{1}{\gamma}\right) dr = \infty,$$

then electrical flow does penetrate closer than ϵ to the origin and no information about γ on $[0, \epsilon)$ can be obtained from Λ_γ.

Our main result on the discrete inverse problem can be roughly stated as the following theorem.

THEOREM 2.16. *A layered conductivity on $D(2m+1, l)$ or $D^*(2m+1, l)$ can be recovered from the corresponding Dirichlet-to-Neumann map if and only if $l \leq m$.*

(See section 4.3 for a refined version.) Theorem 2.16 follows from the general theory in [3] and [4]. In this paper the proof of the uniqueness and the conductivity recovery algorithm are much simpler due to the assumed form of the conductivity.

Our algorithm shows an intimate connection between the discrete inverse problem and the Pick–Nevalinna interpolation problem.

2.10. The case of a half-plane. One often considers the Dirichlet-to-Neumann maps of the lower half-plane with a conductivity that is constant on horizontal lines. For that layered case the following results hold. Let $\overline{L}_{half\text{-}plane}$ be defined by analogy with $\overline{L}$.

THEOREM 2.17.

$$\overline{L}_{half\text{-}plane} = \left\{ \sqrt{-\frac{d^2}{dx^2}}\beta\left(\sqrt{-\frac{d^2}{dx^2}}\right) : \beta \in \mathfrak{B} \right\}.$$

THEOREM 2.18. *A distribution K on $\mathbb{R}\times\mathbb{R}$ is the kernel of a map from $\overline{L}_{half\text{-}plane}$ if and only if*

$$K(\phi, \theta) = h(\phi - \theta),$$

where h is a distribution on $\mathbb{R}$ such that $h(s) = h(-s)$,

$$\int_0^\infty h(s)ds = 0, \qquad \int_0^\infty h(s)(\cos s - 1)ds < \infty,$$

and h is positive definite on $(0, \infty)$.

3. Continuous problem. Our assumption that γ depends only on r makes it possible to reduce our subject of study to a 1-dimensional one.

3.1. Reduction to a 1-dimensional problem.

LEMMA 3.1. *The solution to the Dirichlet problem on $\mathbb{D}$ with the boundary data $e^{ik\theta}, k \in \mathbb{Z}$ is of the form*

$$u_k(r, \theta) = a_k(r)e^{ik\theta}.$$

Proof. Suppose u is γ-harmonic and $u|_{\partial\mathbb{D}} = e^{ik\theta}$. Since the conductivity is constant on circles, for all $\epsilon > 0$

$$v(r, \theta) = u(r, \theta + \epsilon) - u(r, \theta)$$

is also γ-harmonic and $v|_{\partial\mathbb{D}} = e^{ik\theta}(e^{ik\epsilon} - 1)$. Hence, by the uniqueness of the solution of the Dirichlet problem,

$$u_k(r, \theta)(e^{ik\epsilon} - 1) = u_k(r, \theta + \epsilon) - u_k(r, \theta)$$

$$\Rightarrow u_k(r, \theta)e^{ik\epsilon} = u_k(r, \theta + \epsilon)$$

$$\Rightarrow u_k(r, \theta) = a_k(r)e^{ik\theta}. \qquad \square$$

COROLLARY 3.2. *Λ_γ and $\frac{d^2}{d\theta^2}$ have the same eigenfunctions*

$$e^{ik\theta}, k \in \mathbb{Z}.$$

We have that for $k \in \mathbb{N}$

$$\begin{cases} a_k(0) = 0, \\ a_k(1) = 1. \end{cases}$$

Writing (2.1) in polar coordinates gives

$$\frac{d}{dr}\gamma(r) r \frac{d}{dr} a_k(r) - k^2 \frac{\gamma(r)}{r} a_k(r) = 0. \tag{3.1}$$

The eigenvalue of Λ_γ corresponding to $e^{ik\theta}$ and $e^{-ik\theta}$ is

$$R(k) = \gamma \frac{da_k}{dr}(1).$$

For $\gamma \in C$ we make the change of variable

$$x = \int_r^1 \frac{dt}{t\gamma(t)}.$$

Let $x_\infty = \int_0^1 \dfrac{dt}{t\gamma(t)} \le \infty$. We have

$$\left(\frac{1}{\gamma(x)^2}\frac{d}{dx}\right)\frac{d}{dx} a_k(x) = k^2 a_k(x), x \in (0, x_\infty), \tag{3.2}$$

$$\begin{cases} a_k(0) = 1, \\ a_k(x_\infty) = 0, \end{cases}$$

and

$$R(k) = -\frac{da_k}{dx}(0).$$

The investigations by Krein and Kac, outlined in the next section, show that the admittance function R is well-defined even if the operator $\frac{1}{\gamma(x)^2}\frac{d}{dx}$ in (3.2) is replaced by $\frac{d}{dm(x)}$ where $m(x)$ is a distribution function of a positive measure. If $m(x)$ is differentiable then

$$m(x) = \int_0^x \gamma(x)^2 dx = \int_r^1 \frac{\gamma(r) dr}{r}.$$

3.2. Small vibrations of strings. We will give now, without proofs, an outline of some results of Krein and Kac; see also [5].

DEFINITION 3.3. *A* string *is a pair lm, where l is the length of the string* ($0 < l \leq \infty$) *and*

$$m = m(x), x \in [0, l]$$

is a nondecreasing function with

$$0 \leq m(x) < \infty \text{ for } 0 \leq x < l.$$

The value of m at x represents the mass of the interval $[0, x]$.

We note that this definition is slightly different from the one in [9] and [5]. Instead of considering different ways of attaching the right end of the string we allow a weightless interval at that end.

If the right end l of the string is fixed, and a pulsating force

$$F = A \sin \sqrt{\zeta} t, \zeta \notin \mathbb{R}$$

is applied to the left end in the direction perpendicular to the x-axis, the forced oscillation of the left end satisfies the law

$$y = \Omega(\zeta) A \sin \sqrt{\zeta} t.$$

The function Ω is called the *coefficient of dynamic compliance* of the string.

The amplitude function of the oscillation satisfies the following integral equation:

$$\psi(x, \zeta) = \psi(0, \zeta) + \psi'_{-}(0, \zeta)x - \zeta \int_0^x (x - s)\psi(s, \zeta) dm(s). \tag{3.3}$$

If m has the density $\rho(x) = dm/dx$ this equation has an equivalent differential form

$$\frac{1}{\rho(x)} \frac{d^2}{dx^2} \psi(x, \zeta) = -\zeta \psi(x, \zeta). \tag{3.4}$$

The integral form makes the general characterizations below possible.

Let $\phi(x, \zeta)$ and $\theta(x, \zeta)$ be the solutions of (3.3) with the boundary conditions

$$\begin{cases} \phi(0, \zeta) = 1, \\ \dfrac{d\phi}{dx}(0, \zeta) = 0 \end{cases} \quad \text{and} \quad \begin{cases} \theta(0, \zeta) = 0, \\ \dfrac{d\theta}{dx}(0, \zeta) = 1. \end{cases}$$

For every $x \in [0, x_\infty)$ the functions $\phi(x, \zeta)$ and $\theta(x, \zeta)$ are entire functions of ζ. The coefficient of dynamic compliance is determined by these fundamental solutions:

$$\Omega(\zeta) = \lim_{x \to l} \frac{\theta(x, \zeta)}{\phi(x, \zeta)}.$$

Note that

$$\psi(x, \zeta) = \phi(x, \zeta) - \frac{1}{\Omega(\zeta)} \theta(x, \zeta)$$

is the solution of (3.3) with

$$\begin{cases}\psi(0,\zeta)=1,\\ \psi(l,\zeta)=0.\end{cases}$$

The following fundamental theorem is proved in [9]; see also [5].

THEOREM 3.4. *For every function of the form*

$$\Omega(\zeta)=C-\frac{c}{\zeta}+\int_0^\infty\frac{(1+t^2)d\sigma(t)}{t^2-\zeta},\zeta\in\mathbb{C}-[0,+\infty),$$

where σ is a positive measure of bounded variation on $(0,\infty)$ there exists a unique string for which Ω serves as the coefficient of dynamic compliance. And for every string its coefficient of dynamic compliance is of this form.

The measure σ is essentially the spectral measure of the operator $\frac{1}{\rho(x)}\frac{d^2}{dx^2}$. In particular for any $x,y\in[0,x_\infty)$

$$\int_0^\infty\phi(x,\zeta)\phi(y,\zeta)(1+\zeta^2)d\sigma(\zeta)=\delta(x-y).$$

The map from the strings to their coefficients of dynamic compliance is continuous in a sense that if

$$\lim_{n\to\infty}m_n(x)\to m(x)$$

for all $x\in(0,x_\infty)$ such that x is not a jump of m then

$$\lim_{n\to\infty}\Omega_n(\zeta)\to\Omega(\zeta)\text{ for all }\zeta\in\mathbb{C}-[0,+\infty).$$

COROLLARY 3.5. *The formula*

$$\beta(\zeta)=\zeta\Omega(-\zeta^2)$$

gives a one-to-one correspondence between coefficients of dynamic compliance of strings and analytic functions

$$\beta:\mathbb{C}^+\to\mathbb{C}^+$$

with $\beta(\zeta)>0$ for $\zeta>0$.

Proof. Herglotz's theorem in the introduction of this paper. □

3.3. Corollaries of the results of Krein and Kac. We now can put several pieces together to get Theorem 2.6. From section 3.1 and Corollary 3.5 we have that

$$L\subset\left\{\sqrt{-\frac{d^2}{d\theta^2}}\beta\left(\sqrt{-\frac{d^2}{d\theta^2}}\right):\beta\in\mathfrak{B}\right\}.$$

Suppose a sequence $\{\beta_n\}\subset\mathfrak{B}$ is such that $\lim_{n\to\infty}\beta_n(k)$ exists for all $k\in\mathbb{N}$. Then (see [12]) there exists $\beta\in\mathfrak{B}$ such that $\lim_{n\to\infty}\beta_n(k)=\beta(k)$ for all $k\in\mathbb{N}$. Therefore, from the definition of convergence of Dirichlet-to-Neumann maps from section 2.4,

$$\overline{L}\subset\left\{\sqrt{-\frac{d^2}{d\theta^2}}\beta\left(\sqrt{-\frac{d^2}{d\theta^2}}\right):\beta\in\mathfrak{B}\right\}.$$

The map from strings to their coefficients of dynamic compliance is continuous in the topology described in section 3.2. The strings with differentiable mass, which correspond to conductivities from C after the change of variable (section 3.1), are obviously dense in the same topology in the set of all strings. This shows that the containment above is in fact the equality and finishes the proof of Theorem 2.6.

We will now show that the uniqueness in Theorem 3.4 implies Theorem 2.15.

For a conductivity γ in C we put in correspondence the string lm with the length

$$l = \int_0^1 \frac{dr}{r\gamma(r)} \leq \infty,$$

and the mass density

$$\rho\left(\int_r^1 \frac{dt}{t\gamma(t)}\right) = \gamma^2(r).$$

The admittance function corresponding to γ and the coefficient of dynamic compliance of the string are connected by

$$R(\lambda) = \frac{1}{\Omega(-\lambda^2)}. \tag{3.5}$$

By Corollary 3.5 we have that

$$\frac{R(\lambda)}{\lambda} : \mathbb{C}^+ \to \mathbb{C}^+$$

is analytic in $\mathbb{C}^+$ and, therefore, is determined by its values at integers (see [15]). Therefore, the Dirichlet-to-Neumann map Λ_γ determines the coefficient of dynamic compliance $\Omega(\lambda)$ of the correspondent string. Theorem 3.4 guarantees a unique corresponding string with density ρ. The conductivity can be found then by

$$\gamma\left(e^{-\int_0^x \sqrt{\rho(t)}dt}\right) = \sqrt{\rho(x)}. \tag{3.6}$$

3.4. Characterization of kernels: Positive definite functions. We will now explore the role of positive definite functions.

LEMMA 3.6. *Let*

$$f(s) = -\int_0^\infty t\frac{e^{-st} + e^{(s-2\pi)t}}{1 - e^{-2\pi t}}(1 + t^2)d\sigma(t), \qquad s \in (0, \pi), \tag{3.7}$$

where σ is a positive measure of bounded variation on $(0, \infty)$; then

$$\int_0^\pi f(s)(\cos\lambda s - 1)ds = \int_0^\infty \frac{\lambda^2(1 + t^2)d\sigma(t)}{\lambda^2 + t^2}$$

for $\lambda \in \mathbb{C}^+$.

Proof. Tonelli's theorem. □

We are now in a position to prove Theorem 2.14.

Proof. The fact that the kernel of a map from $\overline{L}$ has the properties of the Theorem 2.14 follows directly from Herglotz's theorem and Lemma 3.6. We now show the other direction.

By Bochner's theorem in section 2, if $-f$ is positive definite on $(0, 2\pi)$, there exists a σ-finite measure ν on $\mathbb{R}$ such that

$$f(s) = -\int_{-\infty}^{+\infty} e^{-st} d\nu(t), s \in (0, 2\pi).$$

Since

$$f(s) = f(2\pi - s),$$

there exists a σ-finite measure τ on $(0, \infty)$ such that

$$f(s) = -\frac{1}{2}\int_{-\infty}^{+\infty} e^{-st} + e^{-(2\pi-s)t} d\nu(t) = -\int_0^{\infty} e^{-st} + e^{-(2\pi-s)t} d\tau(t).$$

Since

$$\int_0^{\pi} f(s)(\cos s - 1) ds < \infty,$$

by Lemma 3.6 with $\lambda = 1$ and $d\tau(t) = \frac{t(1+t^2)}{1-e^{-2\pi t}} d\sigma(t)$

$$\int_0^{\infty} \frac{1 - e^{-2\pi t}}{t(1+t^2)} d\tau(t) < \infty$$

$$\Rightarrow f(s) = -\int_0^{\infty} t\frac{e^{-st} + e^{(s-2\pi)t}}{1 - e^{-2\pi t}}(1+t^2) d\sigma(t), \qquad s \in (0, \pi)$$

for a positive measure of bounded total variation σ. Invoking of Lemma 3.6 and Theorem 3.4 finishes the proof. $\square$

The arguments above can be easily transformed to give Theorems 2.17 and 2.18 for the half-plane. Lemma 3.6 should be replaced by the following lemma.

LEMMA 3.7. *Let*

$$g(s) = -\int_0^{\infty} te^{-st}(1+t^2) d\sigma(t), \qquad s \in (0, \infty), \tag{3.8}$$

where σ is a positive measure of bounded variation on $[0, \infty)$; then

$$\int_0^{\infty} g(s)(\cos \lambda s - 1) ds = \int_0^{\infty} \frac{\lambda^2(1+t^2) d\sigma(t)}{\lambda^2 + t^2}$$

for $\lambda \geq 0$.

4. Discrete problem. We will proceed in a manner similar to the continuous case.

4.1. Reduction to a 1-dimensional problem.

LEMMA 4.1. *The solution to the Dirichlet problem on D_n with the boundary data $e^{ik\theta}|_{\partial_n}$ is of the form*

$$u_k(r, \theta) = a_k(r) e^{ik\theta}.$$

Proof. Suppose u is γ-harmonic and $u|_{\partial_n} = e^{ik\theta}$. Since the conductivity is constant on layers, the function

$$v(r,\theta) = u\left(r, \theta + \frac{2\pi}{n}\right) - u(r,\theta)$$

is also γ-harmonic and $v|_{\partial_n} = e^{ik\theta}(e^{i\frac{2\pi k}{n}} - 1)$. Hence, by the uniqueness of the solution of the Dirichlet problem,

$$u_k(r,\theta)(e^{i\frac{2\pi k}{n}} - 1) = u_k\left(r, \theta + \frac{2\pi}{n}\right) - u_k(r,\theta)$$

$$\Rightarrow u_k(r,\theta)e^{i\frac{2\pi}{n}} = u_k\left(r, \theta + \frac{2\pi}{n}\right)$$

$$\Rightarrow u_k(r,\theta) = a_k(r)e^{ik\theta}. \qquad \square$$

COROLLARY 4.2. *Λ_γ and $[\frac{d^2}{d\theta^2}]$ have the same eigenvectors.*

We will now derive an explicit formula for the eigenvalues of Λ_γ in terms of γ. We will do it in a way that emphasizes the relevance of the Sturm–Liouville and beads-on-a-string inverse problems [7] to discrete impedance tomography. Let us first consider the case of $D(n,l)$ with an odd l.

Let $\{\delta_1, \xi_1, \delta_2, \xi_2, \dots, \delta_{\frac{l+1}{2}}\}$ denote the conductivities on layers of $D(n,l)$ starting from the origin. For $k \neq 0$

$$\begin{cases} a_k(0) = 0, \\ a_k(1) = 1, \\ \delta_j(a_k(r_j) - a_k(r_{j-1})) + \delta_{j+1}(a_k(r_j) - a_k(r_{j+1})) + \xi_j a_k(\omega_k^{(n)})^2 = 0. \end{cases}$$

Let $P(\lambda, r_j)$ be the unique solution of the following problem:

$$\begin{cases} P(\lambda, 0) = 0, \\ P(\lambda, r_1) = 1, \\ \delta_j(P(\lambda, r_j) - P(\lambda, r_{j-1})) + \delta_{j+1}(P(\lambda, r_j) - P(\lambda, r_{j+1})) + \lambda^2 \xi_j P(\lambda, r_j) = 0. \end{cases}$$

Let

$$Q(\lambda, r_j) = \delta_j(P(\lambda, r_j) - P(\lambda, r_{j-1})).$$

Then

$$\lambda_k^{(n)} = \frac{Q(\omega_k^{(n)}, 1)}{P(\omega_k^{(n)}, 1)}.$$

We also have

$$\begin{cases} P(\lambda, r_j) = P(\lambda, r_{j-1}) + \frac{1}{\delta_j} Q(\lambda, r_j), \\ Q(\lambda, r_j) = Q(\lambda, r_{j-1}) + \xi_{j-1}\lambda^2 P(\lambda, r_{j-1}). \end{cases} \tag{4.1}$$

Therefore,

$$\frac{Q(\lambda, r_j)}{P(\lambda, r_j)} = \frac{Q(\lambda, r_j)}{P(\lambda, r_{j-1}) + \frac{1}{\delta_j} Q(\lambda, r_j)}$$

$$= \cfrac{1}{\cfrac{1}{\delta_j} + \cfrac{P(\lambda, r_{j-1})}{Q(\lambda, r_j)}} = \cfrac{1}{\cfrac{1}{\delta_j} + \cfrac{P(\lambda, r_{j-1})}{Q(\lambda, r_{j-1}) + \xi\lambda^2 P(\lambda, r_{j-1})}}$$

$$= \cfrac{1}{\cfrac{1}{\delta_j} + \cfrac{1}{\xi_{j-1}\lambda^2 + \cfrac{Q(\lambda, r_{j-1})}{P(\lambda, r_{j-1})}}}. \tag{4.2}$$

Let

$$R(\lambda) = \cfrac{1}{\cfrac{1}{\delta_{\frac{l+1}{2}}} + \cfrac{1}{\xi_{\frac{l-1}{2}}\lambda^2 + \cdots + \cfrac{1}{\cfrac{1}{\delta_3} + \cfrac{1}{\xi_2\lambda^2 + \cfrac{1}{\cfrac{1}{\delta_2} + \cfrac{1}{\xi_1\lambda^2 + \delta_1}}}}}}. \tag{4.3}$$

Then the eigenvalues $\lambda_k^{(n)}$ of Λ_γ are

$$\lambda_k^{(n)} = R(\omega_k^{(n)}).$$

To get the similar formula for other disks one should make corresponding δ_1 or $\frac{1}{\delta_{\frac{l+1}{2}}}$ or both zero.

4.2. Characterizations of the Dirichlet-to-Neumann maps. We consider the function

$$\beta(\lambda) = \frac{1}{\lambda} R(\lambda) = \cfrac{1}{\cfrac{1}{\delta_{\frac{l+1}{2}}}\lambda + \cfrac{1}{\xi_{\frac{l-1}{2}}\lambda + \cdots + \cfrac{1}{\cfrac{1}{\delta_3}\lambda + \cfrac{1}{\xi_2\lambda + \cfrac{1}{\cfrac{1}{\delta_2}\lambda + \cfrac{1}{\xi_1\lambda + \cfrac{1}{\cfrac{1}{\delta_1}\lambda}}}}}}}. \tag{4.4}$$

The function β has the following properties:

1. β is rational,
2. $\beta(\lambda) : \mathbb{C}^+ \to \mathbb{C}^+$,
3. $\beta(\lambda) > 0$ for $\lambda > 0$,
4. $\beta(-\bar{\lambda}) = -\overline{\beta(\lambda)}$.

It turns out that these four properties characterize the continued fractions of the form (4.4); see [11].

Corollary 4.3. *The set of the discrete Dirichlet-to-Neumann maps belongs to*

$$\left\{ \sqrt{-\left[\frac{d^2}{d\theta^2}\right]}\beta\left(\sqrt{-\left[\frac{d^2}{d\theta^2}\right]}\right) : \beta \in \mathfrak{B} \right\}.$$

Let

$$\tau(z) = \frac{1-z}{1+z} : \mathbb{C}^+ \xrightarrow{1-1} \mathbb{D}. \tag{4.5}$$

The characterization in [11] can be restated as the following theorem.

Theorem 4.4. *Let*

$$B : \mathbb{D} \to \mathbb{D}$$

be a real Blaschke product. Then

$$\tau \circ B \circ \tau : \mathbb{C}^+ \to \mathbb{C}^+ \tag{4.6}$$

can be written in unique way as a continued fraction of the form (4.4) *with positive* δ_k, ξ_k. *The number of coefficients in this continued fraction is equal to the number of terms in the product.*

Conversely, every continued fraction of the form (4.4) *with positive* δ_k, ξ_k *can be written in the form* (4.6) *for some real Blaschke product* B.

The following theorem is a consequence of the Pick–Nevanlinna interpolation algorithm (see [6]; see also [12] and [16]).

Theorem 4.5. *Consider*

$$\{z_i\}, \{w_i\} \subset \mathbb{R}^+, i = 1, \dots, m.$$

There exists an analytic function

$$F : \mathbb{C}^+ \to \mathbb{C}^+,$$

$$F(z_i) = w_i, i = 1, \dots, m$$

if and only if the matrix

$$W = \left(\frac{w_i + w_j}{z_i + z_j}\right)_{i,j=1}^{m}$$

is positive semidefinite, if and only if there exists a real Blaschke product B *such that*

$$\tau \circ B \circ \tau(z_i) = w_i, i = 1, \dots, m,$$

which may not be equal to F, e.g., $F \equiv 1$.

If W is singular, the Blaschke product is unique, and the number of terms in it is equal to the size of the largest nonsingular principal minor of W.

If W is not singular there are exactly two desired Blaschke products with the number of terms m, and an infinite family of the Blaschke products with the number of terms $> m$.

Corollary 4.6. *The set of the discrete Dirichlet-to-Neumann maps contains*

$$\left\{ \sqrt{-\left[\frac{d^2}{d\theta^2}\right]}\beta\left(\sqrt{-\left[\frac{d^2}{d\theta^2}\right]}\right) : \beta \in \mathfrak{B} \right\}.$$

This finishes the proof of Theorem 2.5.

4.3. Solution of the discrete inverse problem. Let Λ be an $n \times n$, $n = 2m+1$ discrete layered Dirichlet-to-Neumann map with the nonzero eigenvalues

$$\lambda_k^{(n)}, k = 1, 2, \ldots, m.$$

Consider

$$W = \left(\frac{\lambda_i^{(n)}/\omega_i^{(n)} + \lambda_j^{(n)}/\omega_j^{(n)}}{\omega_i^{(n)} + \omega_j^{(n)}} \right)_{i,j=1}^{m}.$$

If W is singular, there is unique discrete disk $D = D(n, l)$ or $D^*(n, l)$ with unique radially symmetric conductivity γ on it, such that

$$\Lambda(D_\gamma) = \Lambda.$$

Theorems 4.4 and 4.5 give an explicit construction of $D_\gamma(n, l)$. Also, it follows that l is equal to the size of the largest nonsingular principal minor of W, in particular $l < m$.

If W is nonsingular, there are unique conductivities γ, γ' on the disks $D(n, m)$, $D^*(n, m)$, with

$$\Lambda(D_\gamma(n, m)) = \Lambda(D^*_{\gamma'}(n, m)) = \Lambda.$$

For every $D = D(n, l)$ or $D^*(n, l)$ with $l > m$ there are infinitely many condutivities γ with

$$\Lambda(D_\gamma) = \Lambda.$$

REFERENCES

[1] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space*, Dover, New York, 1993.

[2] Y. Colin de Verdiere, *Reseaux Electriques Planaires*, Prepublication de l'Institut Fourier, 225 (1992), pp. 1–20.

[3] E. Curtis, D. Ingerman, and J. Morrow, *Circular planar graphs and resistor networks*, Linear Algebra Appl., 283 (1998), pp. 115–150.

[4] E. Curtis, E. Mooers, and J. Morrow, *Finding conductors in circular networks from boundary measurments*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 781–813.

[5] H. Dym and H. P. McKean, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem*, Academic Press, New York, London, 1976.

[6] J. B. Garnett, *Bounded Analytic Functions*, Academic Press, New York, London, 1981.

[7] F. R. Gantmakher and M. G. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, Gos. izd-vo tekhniko-teoret. lit-ry, Moscow, 1950.

[8] D. Ingerman and J. Morrow, *On a Characterization of the kernel of the Dirichlet-to-Neumann map for a planar region*, SIAM J. Math. Anal., 29 (1998), pp. 106–115.

[9] I. S. Kac and M. G. Krein, *On the Spectral Function of the String*, Amer. Math. Soc. Transl. Ser. 2 103, AMS, Providence, RI, 1974.

[10] S. Karlin, *Total Positivity*, Stanford University Press, Stanford, CA, 1968.

[11] N. Levinson and R. M. Redheffer, *Complex Variables*, Holden-Day, Inc., San Francisco, Cambridge, Amsterdam, 1970.

[12] D. Marshall, *An elementary proof of the Pick-Nevalinna interpolation theorem*, Michigan Math. J., 21 (1974), pp. 219–233.

[13] A. I. Nachman, *Global uniqueness for a two-dimensional inverse boundary value problem*, Ann. of Math. (2), 143 (1996), pp. 71–96.

[14] R. Parker, *The inverse problem of electromagnetic induction: Existence and construction of solutions based on incomplete data*, J. of Geophysical Research, 85 (1980), pp. 4421–4428.

[15] W. Rudin, *Complex and Real Analysis*, McGraw–Hill, New York, 1987.

[16] J. Sylvester, *A convergent layer stripping algorithm for the radially symmetric impedance tomography problem*, Comm. Partial Differential Equations, 17 (1992), pp. 1955–1994.

[17] J. Sylvester and G. Uhlmann, *Inverse boundary value problems at the boundary-continuous dependence*, Comm. Pure Appl. Math., 41 (1988), pp. 197–219.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1235–1269

KAM THEORY AND A PARTIAL JUSTIFICATION OF GREENE'S CRITERION FOR NONTWIST MAPS*

AMADEU DELSHAMS† AND RAFAEL DE LA LLAVE‡

Abstract. We consider perturbations of integrable, area preserving nontwist maps of the annulus (those are maps in which the twist condition changes sign). These maps appear in a variety of applications, notably transport in atmospheric Rossby waves.

We show in suitable two-parameter families the persistence of critical circles (invariant circles whose rotation number is the maximum of all the rotation numbers of points in the map) with Diophantine rotation number. The parameter values with critical circles of frequency ω_0 lie on a one-dimensional analytic curve.

Furthermore, we show a partial justification of Greene's criterion: If analytic critical curves with Diophantine rotation number ω_0 exist, the residue of periodic orbits (that is, one fourth of the trace of the derivative of the return map minus 2) with rotation number converging to ω_0 converges to zero exponentially fast. We also show that if analytic curves exist, there should be periodic orbits approximating them and indicate how to compute them.

These results justify, in particular, conjectures put forward on the basis of numerical evidence in [D. del Castillo-Negrete, J.M. Greene, and P.J. Morrison, *Phys. D.*, 91 (1996), pp. 1–23]. The proof of both results relies on the successive application of an iterative lemma which is valid also for $2d$-dimensional exact symplectic diffeomorphisms. The proof of this iterative lemma is based on the deformation method of singularity theory.

Key words. KAM theory, nontwist maps, periodic orbits, Lagrangian chaos, Rossby waves

AMS subject classifications. 58F05, 58F36, 70K50,76U05,86A99

PII. S003614109834908X

1. Introduction.

1.1. The motivation.

The main goal of this paper is to provide rigorous proofs of several phenomena discovered empirically by del Castillo-Negrete, Greene, and Morrison in [CGM1]. Even if our results will apply to a more general class of maps—see Definitions 1.3, 1.4, etc., for more precise definitions—we will start by describing the results of that paper and the applications of the results we present here.

In [CGM1], the authors consider the two-parameter family of area preserving maps, called there the "quadratic standard map"

$$T_{\omega,\varepsilon}(p,q) = \big(p + \varepsilon \sin(2\pi q), q - (p + \varepsilon \sin 2\pi q)^2 + \omega \pmod 1\big)\,. \tag{1.1}$$

One motivation for such study is that qualitatively similar maps appear naturally in the study of geostrophic flows and indeed in many problems in hydrodynamics and in other applications, mentioned briefly later.

The "unperturbed" map $T_{\omega,0}$

$$T_{\omega,0}(p,q) = (p, q + \Gamma(\omega,p) \pmod 1)\,, \qquad \Gamma(\omega,p) = \omega - p^2 \tag{1.2}$$

*Received by the editors December 11, 1998; accepted for publication (in revised form) October 7, 1999; published electronically May 11, 2000. This work has been partially supported by NATO grant CRG950273.

http://www.siam.org/journals/sima/31-6/34908.html

†Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (amadeu@ma1.upc.es). The research of this author was supported by EC grant ERBCHRXCT-940460, the Spanish grant DGICYT PB94-0215, and the Catalan grant CIRIT 1998SGR–00041.

‡Department of Mathematics, University of Texas, Austin, TX 78712-1082 (llave@math.utexas.edu). The research of this author was supported by NSF grants.

describes a situation where particles in a fluid are moving in a laminar flow whose velocity is faster in the middle ($p = 0$) but slower as we move away from the center of the stream. This is a very common situation in fluid motion, where often the motion slows down as we move closer to edges of the stream. In many applications, it is natural to consider q as an angle. For example, in the description of the jet stream, q corresponds to the longitude and p is a range of latitudes.

The map $T_{\omega,0}$ is an integrable map, since all the circles with fixed p are invariant under $T_{\omega,0}$ and the motion in them is a rigid rotation with rotation number $\Gamma(\omega, p)$. The quantity $\partial\Gamma/\partial p$—usually called the twist—measures the anisochronicity—i.e., the rate of change of frequencies among different invariant circles. The condition $\partial\Gamma/\partial p \neq 0$ is called the "twist condition," and a map which satisfies the twist condition is called a (monotone) twist map. This twist condition does not hold in any map $T_{\omega,0}$ given in (1.2), since $\partial\Gamma/\partial p$ changes sign in $p = 0$. Accordingly, $T_{\omega,0}$ is called a nontwist map. Note that changing of sign is stronger than the twist vanishing in some circle but being otherwise positive. These are the small twist maps, which also appear in many applications. The relevance of the twist condition comes from the celebrated KAM theorem which establishes that the invariant circles whose frequency satisfies a Diophantine condition persist under a small enough—in a smooth norm—area preserving perturbation with zero mean flux. That is to say, twist mappings under perturbation look integrable for a large area. Nontwist maps, on the other hand, experience new phenomena in the area where the twist changes sign. (See later in this introduction for more references.)

The extra term modified by the small parameter ε is representative of the maps that arise when one considers the physical effect of a small periodic oscillation transverse to the channel flow. Such phenomena occur frequently in hydrodynamics when channel flows are destabilized through a Hopf bifurcation. This happens in jet flows in the atmosphere due to Rossby waves. We refer to [C] and [CM] for a detailed description of the fluid mechanics motivation of such models, in particular for the justification of the use of a two-dimensional approximation. In this interpretation, the existence of invariant circles is very important, since they are complete barriers for the mixing of the material in the pole—one of the edges of the latitude p—with the material near the equator—the other edge of p. In the particular model for the atmosphere, these barriers give rise to the creation of "ozone holes" since they isolate the ozone created in the tropics from the regions near the poles.

For area preserving perturbations of twist maps, the twist theorem (see [He] for a quantitative version and [BHS] for an exhaustive description of KAM theory), ensures the persistence, for $|\varepsilon|$ small enough, of those invariant curves with a Diophantine rotation number ω_0:

$$\exists C > 0, \theta \geq 0 : |k \cdot \omega_0 - m|^{-1} \leq C|k|^{\theta-1} \ \forall \ k \in \mathbb{Z}, m \in \mathbb{Z} \setminus \{0\}. \tag{1.3}$$

The set of Diophantine numbers has full measure. A paradigmatic example is $(\sqrt{5} - 1)/2$, which satisfies the inequalities above for $\theta = 0$.

Unfortunately, given a Diophantine rotation number ω_0, the twist theorem cannot be applied to the map $T_{\omega_0,\varepsilon}$ close to the invariant circle $p = 0$, since the twist condition breaks, and moreover the associated rotation number ω_0 lies on the boundary of the range of the rotation numbers $\Gamma(\omega_0, p)$. The paper [CGM1] finds numerically—among other results—numerical evidence for the following claim.

Claim 1.1. *Let $\omega_0 = (\sqrt{5} - 1)/2$. Then, for $|\varepsilon| \ll 1$ there is a smooth curve $\omega(\varepsilon)$ with $\omega(0) = \omega_0$ such that*

(a) *if $\omega > \omega(\varepsilon)$, then $T_{\omega,\varepsilon}$ admits two invariant circles with rotation number ω_0,*
(b) *If $\omega < \omega(\varepsilon)$, then $T_{\omega,\varepsilon}$ admits no invariant circles with rotation number ω_0,*
(c) *If $\omega = \omega(\varepsilon)$, then $T_{\omega(\varepsilon),\varepsilon}$ admits an invariant circle with rotation number ω_0.*

The circle in (c), moreover, is "critical," that is, there exists a change of variables $(p, q) \to (A, \varphi)$ in its neighborhood in such a way that

$$h^{-1} \circ T_{\omega(\varepsilon),\varepsilon} \circ h(A, \varphi) = (A, \varphi + \omega_0 + \kappa A^2) + O(A^3), \qquad \kappa \neq 0$$

(in fact, $\kappa < 0$ for the example in (1.1)).

It is worth noticing that the method used in [CGM1] to assess the existence of the invariant circles is the Greene's criterion, introduced in [Gr]. This criterion asserts that there exists an invariant circle with rotation number ω_0 if and only if

$$\mathrm{Res}(O_{m,n}) := \frac{1}{4}\left[\mathrm{tr}\left(DT^n_{\omega,\varepsilon}\left(O_{m,n}\right)\right) - 2\right] \underset{m/n\to\omega_0}{\longrightarrow} 0$$

for any sequence of periodic orbits O_{mn} of type m/n converging to ω_0.

For the maps $T_{\omega,\varepsilon}$ as in (1.1), the Greene's criterion can be implemented numerically very efficiently. These maps are reversible and, for reversible maps, the search for periodic orbits of type m/n (those are n-periodic orbits which make m complete turns in the angle variable q) in some symmetry lines—not all of the map—can be reduced to finding zeros of one-dimensional functions, a tractable numerical task. In the paper [CGM1] the authors succeed in implementing this criterion, and therefore they also find numerical evidence for the following claim.

Claim 1.2. *Greene's criterion applies.*

In this paper, we will prove rigorous results that justify the experimental results we stated in detail above. We will state and prove a result that justifies Claim 1.1 and another one that justifies one of the implications in Claim 1.2, namely that if there exists an invariant circle, the residue goes to zero.

To our knowledge, the converse—that is, if the residue goes to zero for any sequence of periodic orbits O_{mn} of type m/n converging to ω_0, one can find an invariant circle with rotation number ω_0—remains an open problem even for twist maps. However, we call attention to the work of [KO], which proves that if there are periodic orbits of twist maps which are, in a precise sense, well distributed, one can find an invariant circle with rotation number related to that of the periodic orbit. We also note that if the renormalization group picture can be justified, at least to a certain extent, the Greene's criterion will also be justified and indeed several improvements on that give precise asymptotics of the residue (see [McK]).

It is worth remarking that an easy argument, which we will detail later in Proposition 4.4, shows that if there is a critical invariant circle as above, indeed it is approximated by periodic orbits of type m/n with m/n converging to ω_0. Hence, this criterion is rather effective.

The general theory we will develop will not depend on the exact form for the map but on qualitative features that can be verified in the realistic models. Of course, the map (1.1) is a concrete model introduced for the purpose of discovering qualitative features through numerical calculations.

We also point out that other models having nontwist maps have appeared with other motivations. For example, they appear in celestial mechanics in problems such as the "critical inclination" [K] and in the study of billiards with a boundary moving periodically in time [KMOP1], [KMOP2] or in the study of the motion of particles in magnetic fields [ZZSUC]. As a matter of fact, since the iterates of a twist map are not, in general, twist maps, we expect that they also appear as descriptions of regions of iterates of twist maps. (See, e.g., [BST], [Si].)

These nontwist maps exhibit a very rich phenomenology that has only now started to be explored. The papers cited above as well as [VG], [HH1], [HH2], [Si], [Ha1], [Ha2] contain descriptions and studies of a wealth of phenomena such as “scaling relations,” “reconnection,” “meandering curves,” etc., that deserve to be investigated further. Notably in [Si], [Ha1], there are studies of new phenomena that happen in higher dimensional nontwist maps. In a very recent paper [DMS], it is shown that a generic unfolding of the tripling bifurcation of a fixed point of an area preserving map gives rise to nontwist maps and therefore critical invariant curves appear.

1.2. The methodology. In this paper, we will develop rigorous techniques that can produce results on two problems of the ones mentioned above: The existence of critical invariant circles and the validity of Greene’s criterion. Needless to say, we hope that the techniques that we develop for this purpose (e.g., finding appropriate normal forms and quantitative error estimates of them in neighborhoods) can eventually be used in the study of some of these other phenomena.

About the method of proof we note that there are two basic methods in KAM theory to prove the persistence of invariant tori of exact symplectic mappings or Hamiltonian flows. One is based on applying successive transformations close to the invariant torus and another one is based on solving functional equations that express invariance. Both methods have complementary advantages. The functional equation method leads to very crisp proofs and they are more natural for numerical implementations. On the other hand, the methods based on transformation theory yield more information about the behavior of the map on a neighborhood of the invariant torus.

Since in this paper we wanted to discuss the partial justification of Greene’s criterion, we certainly needed a method based on the transformation theory and it was natural to use the same method for the proof of the persistence of the invariant tori. In the future, we plan to come back to the functional method, especially in connection with a numerical implementation.

The proof we present here will be based on the deformation method. This method was introduced in the study of singularities of mappings [TL], [Mat] and it is very well suited for the study of equivalence of maps in situations where geometric structures are present [BLW], like families of exact symplectic diffeomorphisms. One can also use it for the regular KAM theorem [Ll]. In our case, the use of the deformation method is very natural since the unknown involves a family of maps.

Note that in this situation we are trying to study the persistence of invariant circles whose frequency is on the boundary of the frequencies that are present on the integrable map. This is in contrast with KAM theory, where the nondegeneracy conditions—the so-called twist condition or the more sophisticated Rüssmann conditions (see [BHS, Chapter 4])—imply that the frequency under study is in the interior of the frequencies of the invariant circles in the integrable case.

Since the frequency we want to study is on the boundary of the frequencies, it is not difficult to consider a perturbation of the integrable case in which there is no invariant circle with the frequency we want. (It suffices to consider an integrable perturbation in which we just add—or subtract—an extra rotation so that all the invariant circles persist, but their rotation number is changed.)

Speaking heuristically, what we will do is to consider the regular perturbation theory supplemented with a choice of $\omega(\varepsilon)$. The regular perturbation theory may force the ω_0 out of the range of frequencies, but we will find the extra rotation $\omega(\varepsilon)$ that puts it on the boundary. Since in this method of proof one needs to consider families all the time, the use of the deformation method seems particularly well justified.

On a more technical level, we note that the proof will be based on an iterative lemma (Lemma 3.6) that describes how it is possible to obtain transformations that reduce the system to integrable. Moreover, we will present bounds on the error of this reduction depending on the domain. This iterative lemma can be applied repeatedly in different ways depending on how one plays the tradeoff between domain loss and accuracy. One can try to make the error decrease very fast at the price that the domain decreases very fast or one can make the error decrease slowly on a larger domain. In this way, one can obtain a unified approach towards KAM theory and towards exponentially small estimates, which we will show justify Greene's criterion. This approach has precedents in [DG1]. Since the iterative lemma, as well as the deformation method are widely applicable, we have developed it in an arbitrary dimension. The geometric considerations that lead to the KAM theorem for critical circles and to the Greene's criterion seem to be different in higher dimensions, so we have postponed the discussion of this part.

1.3. The results.

Now we turn to making all these ideas more precise.

Definition 1.3. *We say that a circle S, invariant under an area preserving map T of $\mathbb{R}\times\mathbb{T}^1 \equiv M$, is a critical invariant circle if there exists a canonical transformation $h : [-\delta,\delta]\times\mathbb{T}^1 \to M$ in such a way that*

$$h^{-1}\circ T\circ h(A,\varphi) = (A, \varphi+\omega_0+\kappa A^2) + O(A^3)$$

with $\kappa \neq 0$ and $h(\{0\}\times\mathbb{T}^1) = S$.

Remark. The definition of a critical circle includes in its hypothesis that the motion on the circle is conjugate to a rotation of ω_0. We will not include the ω_0 in the notation since it will be understood from the context.

We also recall—and we will develop it in more detail later in Lemma 4.2—that there is an analogue of Birkhoff normal form in a neighborhood of an invariant circle with a Diophantine rotation. (In the twist map case, this was also considered in [OS, FL].) Given $N \in \natural$, it is possible to find coefficients $\kappa_1,\dots,\kappa_N$ and a canonical transformation h such that

$$h^{-1}\circ T\circ h(A,\varphi) = (A, \varphi+\omega_0+\kappa_1 A+\kappa_2 A^2+\cdots+\kappa_M A^M) + O(A^{M+1}). \tag{1.4}$$

The coefficients $\kappa_1,\dots,\kappa_N$ are uniquely defined and are properties of the invariant circle. In this language, critical circles are those for which $\kappa_1 = 0$, $\kappa_2 \neq 0$.

Definition 1.4. *We will call an invariant circle nondegenerate when the normal form* (1.4) *does not vanish identically. That is, we can find $M \in \natural$ such that $\kappa_1 = \cdots = \kappa_{M-1} = 0$, $\kappa_M \neq 0$.*

Our result to justify Claim 1.1 is the following theorem.

Theorem 1.5. *Let ω_0 be a Diophantine number as in* (1.3), *$f_{\omega,\varepsilon}$ be a family of mappings from $\mathbb{R}^1\times\mathbb{T}^1$ to itself satisfying*

(i) *$f_{\omega,\varepsilon}(p,q)$ is analytic in*

$$|\omega-\omega_0| < \rho_0\ , \quad |\varepsilon| < \rho_0\ , \quad |\Im\, q| < \beta_0\ , \quad |p| < \rho_0$$

and takes real values for ω,ε,p,q real;

(ii) *$f_{\omega,\varepsilon}$ is exact symplectic $\forall\omega,\varepsilon$,*

(iii) *$f_{\omega,0}(p,q) = (p, q+\Gamma(\omega,p))$ with*

$$\Gamma(\omega_0,0)=\omega_0, \qquad \frac{\partial}{\partial p}\Gamma(\omega_0,0)=0,$$
$$\frac{\partial^2}{\partial p^2}\Gamma(\omega_0,0)=t<0, \qquad \frac{\partial}{\partial \omega}\Gamma(\omega_0,0)=s>0.$$

Then, we can find a $\delta > 0$ and an analytic function ω defined for $|\varepsilon|$ sufficiently small and taking real values for ε real in such a way that

(a) *$f_{\omega(\varepsilon),\varepsilon}$ has exactly one critical invariant circle in $[-\delta,\delta]\times\mathbb{T}^1$;*

(b) *if $\omega < \omega(\varepsilon)$, $f_{\omega,\varepsilon}$ has no points in $[-\delta,\delta]\times\mathbb{T}^1$ with rotation number ω_0, and if $\omega > \omega(\varepsilon)$ there are two invariant circles of $f_{\omega,\varepsilon}$ in $[-\delta,\delta]\times\mathbb{T}^1$ which are not critical.*

Remark. It is possible to change hypothesis (iii) of Theorem 1.5, to be that t is positive. It suffices to change the inequalities between ω, $\omega(\varepsilon)$ in part (b) of the conclusions and the proof goes through without change (similarly if s is negative in (iii)).

The precise meaning in which Greene's criterion can be justified in the following theorem.

Theorem 1.6. *Let $f_{\omega,\varepsilon}$ be an analytic area preserving diffeomorphism of the annulus. Assume that $f_{\omega,\varepsilon}$ admits an analytic invariant circle on which the motion is analytically conjugate to a rotation with Diophantine number ω_0 and which is nondegenerate in the sense of Definition* 1.4.

Then, we can find $C_1, C_2, \mu > 0$ (depending on ω_0, the map, and the torus) such that for any sequence of periodic orbits O_n of type p_n/q_n which are converging to the analytic invariant circle and such that $|\omega_0 - p_n/q_n| \le 1/q_n$, we have

$$\operatorname{Res}(O_n) \le C_1 \exp\left(-C_2|\omega_0 - p_n/q_n|^{-\mu}\right). \tag{1.5}$$

We will also show that there is one such sequence of periodic orbits converging to the nondegenerate circle. Of course, when the circle is critical, depending on the sign of $\omega - p_n/q_n$ we will find either two or four periodic orbits. For more general nondegenerate circles, when M is even we will find two or four periodic orbits of type p_n/q_n depending on the sign of $\omega - p_n/q_n$ and when M is odd we will find two irrespective of the sign or $\omega - p_n/q_n$.

Remark. The proof that the residue goes to zero faster than any power is significantly easier than the proof with an explicit rate.

2. The deformation method.

In this section we recall the basis of the deformation method for symplectic maps. This method was introduced in singularity theory [TL], [Mat], but it was remarked later that it can be used very effectively to obtain structure theorems for volume preserving maps of a manifold [Mo1], or for symplectic maps [W] giving a very direct proof of Darboux theorem. More details and other applications can be found in [LMM], [Ll], [BLW] and in several other places.

In this section, the dimension of the space will not play a role, so we will consider M a $2d$-dimensional manifold.

We recall that a 2-form ϖ on M is a symplectic form if it is closed and has full rank. (Of course, the fact that ϖ has full rank implies that the dimension of M is even; this is why we chose the notation $2d$ for it.) We will be especially interested in the case when ϖ is exact. That is, there exist a 1-form ϑ such that $\varpi = d\vartheta$.

A diffeomorphism f is symplectic when $f_*\varpi = \varpi$. For ϖ exact, this is equivalent to $d(f_*\vartheta - \vartheta) = 0$. We say that a symplectic map f is exact when $f_*\vartheta - \vartheta = dS$ for some function S, called the primitive function of f.

Given a family of diffeomorphisms f_ε, we denote by $\mathcal{F}_\varepsilon$ the vector field defined by

$$\frac{d}{d\varepsilon} f_\varepsilon = \mathcal{F}_\varepsilon \circ f_\varepsilon \tag{2.1}$$

and refer to $\mathcal{F}_\varepsilon$ as the generator of f_ε. Note that a family determines the generator and, conversely, by the uniqueness theorem for ordinary differential equations (ODEs), a family is determined by its initial point f_0 and its generator, when the generator is $\mathcal{C}^1$. (We will always assume that this is the case.)

The main idea of the deformation method is to always work with the generators, which, when the families are differentiable enough so that the uniqueness theorem for ODEs applies, is equivalent to working with the families. When the diffeomorphisms are symplectic, further simplifications are possible. Using Cartan's formula for Lie derivatives and that ϖ is closed we obtain

$$\begin{aligned} \frac{d}{d\varepsilon} f_{\varepsilon *}\varpi &= f_{\varepsilon *}(d(i(\mathcal{F}_\varepsilon)\varpi) + i(\mathcal{F}_\varepsilon)\, d\varpi) = f_{\varepsilon *}(d(i(\mathcal{F}_\varepsilon)\varpi)), \\ \frac{d}{d\varepsilon} f_{\varepsilon *}\vartheta &= d(f_{\varepsilon *}(i(\mathcal{F}_\varepsilon))\vartheta) + f_{\varepsilon *}(i(\mathcal{F}_\varepsilon)\varpi). \end{aligned} \tag{2.2}$$

If f_ε is symplectic, $\frac{d}{d\varepsilon} f_{\varepsilon *}\varpi = 0$, and then we see that

$$d(i(\mathcal{F}_\varepsilon)\varpi) = 0. \tag{2.3}$$

If f_ε is exact symplectic, $d\left(\frac{d}{d\varepsilon} S_\varepsilon\right) - f_{\varepsilon *} d(i(\mathcal{F}_\varepsilon)\varpi) = f_{\varepsilon *}(i(\mathcal{F}_\varepsilon)\vartheta)$ and, therefore,

$$i(\mathcal{F}_\varepsilon)\varpi = dF_\varepsilon \tag{2.4}$$

with $F_\varepsilon = \left(\frac{d}{d\varepsilon} S_\varepsilon\right) \circ f_\varepsilon - i(\mathcal{F}_\varepsilon)\vartheta$.

Conversely, if $\mathcal{F}_\varepsilon$ satisfies (2.3) or (2.4) and f_0 is symplectic or exact symplectic, the family f_ε is symplectic or exact symplectic as can be seen integrating (2.2).

Within this paper, we will refer to F_ε as the Hamiltonian for the family f_ε. Note that given f_ε, (2.4) determines F_ε up to a function of zero differential hence, constant on each connected component of its domain of definition. This justifies calling F_ε "the Hamiltonian" if we think of Hamiltonians as equivalent when they differ in a function with zero differential. This identification is natural since two Hamiltonian differing by a function with zero differential generate the same dynamics.

Conversely, for a $\mathcal{C}^2$ Hamiltonian F_ε, given that ϖ is full rank, (2.4) determines $\mathcal{F}_\varepsilon$, and it is $\mathcal{C}^1$. This $\mathcal{F}_\varepsilon$ and f_0 determine f_ε by the uniqueness result for ODEs.

Hence, for sufficiently smooth families it is equivalent to work with the Hamiltonians and the initial points of the families.

The main idea of the deformation method for exact symplectic maps is to reformulate all the problems in terms of Hamiltonians. As it turns out, the equations involving generators are linear. This is to be expected since we can heuristically think of generators as infinitesimal transformations and all the equations among infinitesimal quantities are linear. Moreover, using Hamiltonians, the otherwise complicated constraint of the transformations being exact symplectic is implemented automatically, and the resulting equations involve only functions. Hence, rather than dealing with nonlinear equations among diffeomorphisms satisfying nonlinear constraints, we just have to deal with a linear equation among functions.

We will follow the convention of denoting families in lowercase f_ε, their generators in calligraphic font $\mathcal{F}_\varepsilon$, and the Hamiltonians in uppercase F_ε.

PROPOSITION 2.1. *Let $f_\varepsilon, g_\varepsilon$ be exact symplectic families and k an exact symplectic diffeomorphism. Then, the Hamiltonian of the families formed out of them are given in the following table.*

Family	Hamiltonian
$f_\varepsilon \circ g_\varepsilon$	$F_\varepsilon + f_{\varepsilon *} G_\varepsilon = F_\varepsilon + G_\varepsilon \circ f_\varepsilon^{-1}$
f_ε^{-1}	$-F_\varepsilon \circ f_\varepsilon$
$g_\varepsilon^{-1} \circ f_\varepsilon \circ g_\varepsilon$	$F_\varepsilon \circ g_\varepsilon - G_\varepsilon \circ g_\varepsilon + G_\varepsilon \circ f_\varepsilon^{-1} \circ g_\varepsilon$
$k^{-1} \circ f_\varepsilon \circ k$	$F_\varepsilon \circ k$
$f_\varepsilon \circ k$	F_ε

The computations needed to work out this table can be found in [LMM], [BLW]. In the latter paper one can find similar tables for volume preserving or contact families.

Since in perturbation theory one does not always have a family of diffeomorphisms but just two diffeomorphisms that are close, it is worth remarking that given two symplectic diffeomorphisms that are close, one can always interpolate them by a family with small Hamiltonian. If the two maps are exact, the family can be chosen to be exact. This is an immediate consequence of the general fact that symplectic (or exact symplectic) maps form a Banach manifold (see [W]). We just sketch a direct construction whose details appear in [BLW]. An alternative, old fashioned proof can be obtained using generating functions. (Interpolate the generating functions.) Unfortunately, since it is impossible to obtain generating functions that are globally defined, one has to also use partitions of unity and fragmentation lemmas and the proof becomes cumbersome.

Given f_0, f_1 symplectic and close enough, we can find a family of diffeomorphisms f_ε interpolating between them (e.g., $f_\varepsilon(x) = \exp_{f_0(x)} \varepsilon \exp_{f_0(x)}^{-1} f_1(x)$ where exp is the Riemannian exponential map). The family f_ε will not be symplectic. In general, $f_{\varepsilon *}\varpi = \varpi_\varepsilon$ where ϖ_ε is a family of symplectic forms. Note that, by our assumptions $\varpi_0 = \varpi_1 = \varpi$. Using Moser's construction [Mo1]—we refer to [LMM], [BLW] for the elementary justification of the smooth dependence on parameters in Moser's construction—we can find h_ε close to the identity in such a way that $h_{\varepsilon *}\varpi_\varepsilon = \varpi$. Moreover, $h_0 = h_1 = \mathrm{Id}$. Then $\tilde{f}_\varepsilon = h_\varepsilon \circ f_\varepsilon$ satisfies $\tilde{f}_0 = f_0$, $\tilde{f}_1 = f_1$, $\tilde{f}_{\varepsilon *}\varpi = \varpi$. If $\varpi = d\vartheta$ then $\varpi_\varepsilon = d\vartheta_\varepsilon$ with $\vartheta_\varepsilon = f_{\varepsilon *}\vartheta$. Also $(h_\varepsilon \circ f_\varepsilon)_* \vartheta - \vartheta$ is closed. It is then possible to choose g_ε close to the identity in such a way that $(g_\varepsilon \circ h_\varepsilon \circ f_\varepsilon)_* \vartheta - \vartheta$ is exact (e.g., on the annulus choose translations in the radial direction and in another manifolds choose a displacement in a neighborhood of paths that generate the homology).

We have, therefore, established the following lemma.

LEMMA 2.2. *Let f_0 be a $\mathcal{C}^\infty$ (resp., $\mathcal{C}^\omega$) symplectic (resp., exact symplectic) diffeomorphism of a manifold.*

If f_1 is a symplectic (resp., exact symplectic) diffeomorphism close to f_0 we can find a $\mathcal{C}^\infty$ (resp., $\mathcal{C}^\omega$) family f_ε of symplectic (resp., exact symplectic) diffeomorphisms interpolating between f_0 and f_1.

Moreover, we can arrange that the generators and therefore the Hamiltonians of the isotopy are arbitrarily small in the $\mathcal{C}^\infty$ (resp., $\mathcal{C}^\omega$) topology by assuming that f_1 is arbitrarily close to f_0.

3. Proof of Theorem 1.5 using the deformation method.

3.1. Heuristic discussion. The proof we present here starts with the observation that the result would be obvious if we had a family of the form

$$i_{\omega,\varepsilon}(p,q) = (p, q + \Omega(\omega,\varepsilon,p)) \tag{3.1}$$

in which the p is conserved and the q is translated by $\Omega(\omega,\varepsilon,p)$, which depends on p and on external parameters and is close to the frequency $\Gamma(\omega,p)$ satisfying hypothesis (iii) of Theorem 1.5. We will refer to such families as integrable.

If we require that the set $p = p_0$ is an invariant circle with rotation ω_0, we obtain the implicit equation

$$\Omega(\omega,\varepsilon,p_0) = \omega_0\ . \tag{3.2}$$

The possibility of finding solutions of (3.2) is described by singularity theory and the phenomenon of a critical invariant circle corresponds to the situation when $\Omega(\omega,\varepsilon,p_0)-\omega_0$ has a fold:

$$\Omega(\omega,\varepsilon,p_0) - \omega_0 = 0, \qquad \partial_p\Omega(\omega,\varepsilon,p_0) = 0.$$

The equation for $\omega(\varepsilon)$ is precisely the equation for the edge of a fold. We will parameterize the folding surface (3.2) as the set of points $(\Upsilon(\varepsilon,p),\varepsilon,p)$ for an appropriate function Υ:

$$\Omega(\omega,\varepsilon,p) = \omega_0 \iff \omega = \Upsilon(\varepsilon,p). \tag{3.3}$$

Then, a critical invariant circle takes place at $p = p_0 = p_0(\varepsilon)$ if $\partial_p\Upsilon(\varepsilon,p_0) = 0$, and $\omega(\varepsilon) = \Upsilon(\varepsilon,p_0(\varepsilon))$.

A standard technique in KAM theory is to make changes of variables so that in the new system of coordinates, the properties of the map are apparent from its expression. In the present case, we try to find g_ε in such a way that

$$\tilde{f}_{\omega,\varepsilon} = g_\varepsilon^{-1} \circ f_\varepsilon \circ g_\varepsilon \tag{3.4}$$

has the desired form (3.1).

Unfortunately, in general it is not possible to obtain a change of variables reducing to (3.1) in the whole phase space. We only know how to do it approximately in a subset of the domain in (ω,ε,p) for which $\Omega(\omega,\varepsilon,p) = \omega_0$.

Hence we will use an iterative scheme in which at step n, the system will be (described in the notation of the deformation method by the initial point of the isotopy and the generating Hamiltonian)

$$f^n_{\omega,0}(p,q) = (p, q+\Gamma(\omega,p)); \qquad F^n_{\omega,\varepsilon}(p,q) = I^n_{\omega,\varepsilon}(p) + E^n_{\omega,\varepsilon}(p,q), \tag{3.5}$$

where $E^n_{\omega,\varepsilon}$ is "small" in a neighborhood of $\{p = 0\}$.

The Hamiltonian $I^n_{\omega,\varepsilon}(p)$ corresponds to a deformation of the form

$$i^n_{\omega,\varepsilon}(p,q) = (p, q + \Omega^n(\omega,\varepsilon,p)), \tag{3.6}$$

where

$$\Omega^n(\omega,\varepsilon,p) = \Gamma(\omega,p) + \int_0^\varepsilon ds \frac{\partial}{\partial p} I^n_{\omega,s}(p) \tag{3.7}$$

when we assume that $i_{\omega,0} = f_{\omega,0}$. Hence, the $I^n_{\omega,\varepsilon}$ should be thought of as the integrable part of the Hamiltonian $F^n_{\omega,\varepsilon}$. We will think of $E^n_{\omega,\varepsilon}$ as an error term that is to be made smaller and smaller in the iterative process.

Remark. We note that the decomposition of a Hamiltonian into an integrable part and an small part is not uniquely defined. A particularly natural one would be to take the integrable part to be the average over the q. Nevertheless, we will not be assuming that this natural decomposition is taken, just that such a decomposition exists.

Remark. Note that when we consider perturbations of an integrable system, we can write the integrable part in Ω and, hence, assume that $I^0_{\omega,\varepsilon}(p) = 0$.

The main ingredient of the proof of Theorem 1.5 will be an algorithm that, given a family as in (3.5), finds a transformation $g^n_{\omega,\varepsilon}$ defined in a neighborhood of the surface $\Omega^n(\varepsilon, \omega, p) = \omega_0$ such that setting $f^{n+1}_{\omega,\varepsilon} = (g^n_{\omega,\varepsilon})^{-1} \circ f^n_{\omega,\varepsilon} \circ g^n_{\omega,\varepsilon}$ we have

$$F^{n+1}_{\omega,\varepsilon}(p,q) = I^{n+1}_{\omega,\varepsilon}(p) + E^{n+1}_{\omega,\varepsilon}(p,q),$$

where $E^{n+1}_{\omega,\varepsilon}$ is much smaller than $E^n_{\omega,\varepsilon}$ and $I^{n+1}_{\omega,\varepsilon}$ differs little from $I^n_{\omega,\varepsilon}$ in a domain which will be chosen appropriately (a smaller neighborhood of the surface $\Omega^{n+1} = \omega_0$).

Since Ω^{n+1} is close to Ω^n, the folding surfaces defined by $\Omega^{n+1} = \omega_0$ and by $\Omega^n = \omega_0$ are very close. Quantitative estimates will show that the $E^n_{\omega,\varepsilon}$'s decrease superexponentially and that the $g^n_{\omega,\varepsilon}$'s differ from the identity by a superexponentially small quantity in neighborhoods of the surfaces $\Omega^{n+1} = \omega_0$. As it turns out, we will have to choose these neighborhoods to become superexponentially thinner. The transformations will be defined in these thin slivers in the ω, ε, p coordinates and in domains in q which include complex extensions of $\mathbb{T}^1$ so that the size of the imaginary extension of the domain remains bounded from below.

Similarly, the functions Ω^n converge to a function Ω^∞. Therefore, the surfaces $\hat{\Omega}^n \equiv (g^1 \circ \cdots \circ g^n)^{-1}\{\Omega^n = \omega_0\}$ converge to a surface $\hat{\Omega}^\infty$. Since each of the surfaces $\{\Omega^n = \omega_0\}$ is foliated by smooth circles invariant by $F \circ g^1 \circ \cdots \circ g^n$ up to superexponentially small errors, it follows that $\hat{\Omega}^\infty$ is foliated by smooth circles invariant by F.

For the benefit of experts, we point out that an alternative method to prove Theorem 1.5 could have been to use the nondegeneracy in ω to prove a KAM theorem for all small enough ε and p. (That is, we fix ε and p, but allow ourselves to choose the ω.) Even if not all methods to prove KAM theorems would have worked, it seems that methods based on the “translated curve method” works since one can use the ω to adjust the frequency. Then, one needs to prove the analytic dependence of the circle on the parameter ε and to prove that there is indeed a fold.

The method we develop in this paper seems more appealing since one has an understanding of the folding surface at all the stages of the iteration and it is certainly not longer to write in all detail.

Moreover, we can use much of the technology developed along these lines, to prove the partial converse of Greene's theorem. In particular, Lemma 3.6 is the crux of the iterative step in the proofs of both problems. The difference between the KAM theorem and the proof of the exponentially small estimates that imply Greene's criterion lies only in different choices on how we iterate the method. In the KAM theorem, we lose domain very fast and drive the errors to zero very fast. In the exponentially small estimates, we reduce the domains more slowly and do not obtain convergence, but the estimates are valid in a larger domain.

We also call attention to the fact that Lemma 3.6 is valid in any dimension. It is only the geometric considerations about domains that one uses to conclude Theorem 1.5 and Theorem 1.6 that require the fact that we are working in an annulus. We think that this restriction can be lifted with some small amount of extra effort.

3.2. Notation and elementary estimates. Since the iterative step will rely on making transformations on functions in such a way that the errors become smaller, we will need to define appropriate norms. We will also need to be able to manipulate sets where our transformations will be defined. (As usual in KAM theory, one has to consider functions defined in decreasing sets). In this section, we collect the definitions of the norms, parameterizations of sets that we will use later as well as some elementary lemmas and propositions dealing with them.

Since Lemma 3.6 is valid in any number of dimensions, we will be considering maps in $\mathbb{R}^d \times \mathbb{T}^d$ till the end of section 3.5.

We recall the standard definition that $\omega_0 \in \mathbb{R}^d$ is said to be Diophantine of exponent θ if we can find a $C > 0$ such that $\forall k \in \mathbb{Z}^d, m \in \mathbb{Z}$ we have

$$|k \cdot \omega_0 - m|^{-1} \leq C|k|^{\theta - 1}. \tag{3.8}$$

This is the definition of Diophantine vectors that appears naturally in KAM theory for maps. (The definition that appears naturally in KAM theory for flows is slightly different.)

Besides the above standard definition, in this paper we will use the following notations.

We will denote by $I_{a,b}$ the real interval $[a,b]$, by $B_{x,c}$ the closed ball in $\mathbb{R}^d$ with center $x \in \mathbb{R}^d$ and radius $c > 0$, and by $\mathbb{T}^d$ the d-dimensional torus $\mathbb{R}^d/\mathbb{Z}^d$.

We will also denote by $I_{a,b,\delta} = \{z \in \mathbb{C} \mid d(z, I_{a,b}) \leq \delta\}$, $B_{x,c,\delta} = \{z \in \mathbb{C}^d \mid d(z, I_{a,b}) \leq \delta\}$. Similarly we will denote by $\mathbb{T}^d_\beta$ the complex extensions on the torus $\mathbb{T}^d$ of a distance β.

Given a set $U = B_{x_1,c_1,\delta} \times I_{a_2,b_2,\delta} \times B_{x_3,b_3,\delta}$ and a function $\Omega : U \to \mathbb{C}^d$, we will denote for $\alpha, \beta > 0$

$$\begin{aligned} \Sigma_{\beta,U} &= \{(\omega, \varepsilon, p, q) \mid (\omega, \varepsilon, p) \in U, |\Im\, q| \leq \beta\} = U \times \mathbb{T}^d_\beta, \\ \Sigma_{\Omega,\alpha,\beta,U} &= \{(\omega, \varepsilon, p, q) \in \Sigma_{\beta,U} \mid |\Omega(\omega, \varepsilon, p) - \omega_0| \leq \alpha\}. \end{aligned} \tag{3.9}$$

The way to think about $\Sigma_{\Omega,\alpha,\beta,U}$ is as the Cartesian product of a thin film—of width α, which will be extremely small in the proof—around a portion of surface given by the equation $\Omega(\omega, \varepsilon, p) = \omega_0$ and a complex extension of width β of the torus. The parameter U just limits which portion of the surface we are considering and it plays a somewhat minor role.

Note that, for the sake of notation, we are suppressing some of the parameters on which $\Sigma_{\Omega,\alpha,\beta,U}$ depends. Notably ω_0. We hope that this does not lead to confusion in the proof since the values of these parameters will be kept fixed. The ω_0 will be that appearing in Theorem 1.5 and, hence, will not change throughout the proof.

We will introduce the notation U_σ to denote a domain formed by restricting the domain only in the variable p by an amount $\sigma > 0$, that is, $U = B_{x_1,c_1,\delta} \times I_{a_2,b_2,\delta} \times B_{x_3,b_3,\delta-\sigma}$.

This will be used later since we need to reduce the domains in phase space (to guarantee that compositions make sense) but the domains in parameters are not affected.

Given a complex domain Σ, we will denote by $\|F\|_\Sigma \equiv \sup_{x \in \Sigma} |F(x)|$ and by χ^Σ the Banach space of functions analytic in Σ (analytic in the interior and continuous up to the boundary) equipped with the norm $\|\cdot\|_\Sigma$. In particular, for $\Sigma = \Sigma_{\beta,U}$, $\Sigma = \Sigma_{\Omega,\alpha,\beta,U}$ of the form (3.9), for typographical reasons, we will write $\|\cdot\|_{\Sigma_{\beta,U}}$ as $\|\cdot\|_{\beta,U}$ and $\|\cdot\|_{\Sigma_{\Omega,\alpha,\beta,U}}$ as $\|\cdot\|_{\Omega,\alpha,\beta,U}$.

For a function $F : U \times \mathbb{T}^d_\beta \to \mathbb{C}$, where $U = B_{x_1,c_1,\delta} \times I_{a_2,b_2,\delta} \times B_{x_3,b_3,\delta}$, we define the partial Fourier expansion

$$F_{\omega,\varepsilon}(p,q) = \sum_{k\in\mathbb{Z}^d} \hat{F}_{\omega,\varepsilon;k}(p) e^{2\pi i\, k\cdot q}.$$

The coefficients are unique in the regularity classes we will be considering.

For this kind of functions depending on parameters, we will use the notation ∇ to denote the derivatives with respect to the variables, not with respect to the parameters. Hence

$$\nabla F_{\omega,\varepsilon}(p,q) = \left(\frac{\partial}{\partial p} F_{\omega,\varepsilon}(p,q), \frac{\partial}{\partial q} F_{\omega,\varepsilon}(p,q) \right).$$

In the cases that we will need to consider derivatives with respect to the parameters, we will write them explicitly.

We recall that the well-known Cauchy inequalities allow us to bound derivatives (in a domain) and Fourier coefficients of a function in terms of its size in a (slightly larger) domain.

Lemma 3.1. *Let $U = B_{x_1,c_1,\delta} \times I_{a_2,b_2,\delta} \times B_{x_3,b_3,\delta}$, $\tilde{U} \subset U$ be a domain that is at a distance $\sigma > 0$ from the complement of U, and $F : U \times \mathbb{T}^d_\beta \to \mathbb{C}$ analytic. Then,*

$$\begin{aligned} \|\nabla^m F\|_{\beta-\sigma,\tilde{U}} &\le K\sigma^{-m} \|F\|_{\beta,\tilde{U}}, \\ \|\partial^m_\omega F\|_{\beta,\tilde{U}}, \|\partial^m_\varepsilon F\|_{\beta,\tilde{U}} &\le K\sigma^{-m} \|F\|_{\beta,\tilde{U}}, \\ |\hat{F}_{\omega,\varepsilon;k}(p)| &\le K e^{-2\pi\beta|k|} \|F\|_{\beta,\{(\omega,\varepsilon,p)\}}. \end{aligned}$$

The well-known proof is based on expressing the Fourier coefficients or derivatives as integrals over paths and deforming them in the complex domain. It can be found in many reference books and we will not reproduce it here.

3.3. The iterative step. In this subsection, we will specify the iterative step of the algorithm and we develop quantitative estimates that will later lead to the possibility of iterating it and showing it converges. Most of these estimates will be used also in Theorem 1.6 on the partial justification of Greene's criterion.

We recall that for the purposes of the iterative lemma, Lemma 3.6, the dimension of the space will be irrelevant, so we will state the results in the $2d$-annulus $\mathbb{R}^d \times \mathbb{T}^d$.

At the beginning of the iterative step, we will be given a family of exact symplectic maps $f_{\omega,\varepsilon}$ defined on a subset of $\mathbb{R}^d \times \mathbb{T}^d$ endowed with the standard symplectic structure.

$$f_{\omega,0}(p,q) = (p, q + \Gamma(\omega,p)), \qquad F_{\omega,\varepsilon}(p,q) = I_{\omega,\varepsilon}(p) + E_{\omega,\varepsilon}(p,q), \tag{3.10}$$

where $F_{\omega,\varepsilon}$, the Hamiltonian of the deformation $f_{\omega,\varepsilon}$, is defined in a set $\Sigma_{\Omega,\alpha,\beta,U}$ of the type described in (3.9), with

$$U = B_{\omega_0,\gamma,\delta} \times I_{[-1,1],\delta} \times B_{0,\gamma,\delta}$$

for some $\gamma > 0$, $0 < \delta < 1$, where ω_0 is a Diophantine vector (e.g., it satisfies (3.8)) and

$$\Omega(\omega, \varepsilon, p) = \Gamma(\omega, p) + \int_0^\varepsilon ds \frac{\partial}{\partial p} I_{\omega,s}(p). \tag{3.11}$$

Since $\Omega(\omega, 0, p) = \Gamma(\omega, p)$, from the hypotheses of Theorem 1.5 we will also assume that Ω is nondegenerate, that is, that we have

$$\left\|(\partial_\omega \Omega)^{-1}\right\|_U \leq A, \qquad \left\|(\partial_p^2 \Omega)^{-1}\right\|_U \leq B. \tag{3.12}$$

The goal of the iterative step is to determine $g_{\omega,\varepsilon}$, $g_{\omega,0} = \mathrm{Id}$, in such a way that $\tilde{f}_\varepsilon = g_{\omega,\varepsilon}^{-1} \circ f_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}$ has Hamiltonian

$$\tilde{F}_{\omega,\varepsilon}(p, q) = \tilde{I}_{\omega,\varepsilon}(p) + \tilde{E}_{\omega,\varepsilon}(p, q), \tag{3.13}$$

where $\tilde{I}_{\omega,\varepsilon}$, $\tilde{E}_{\omega,\varepsilon}$ will be defined in a slightly smaller domain than $I_{\omega,\varepsilon}$, $E_{\omega,\varepsilon}$ and where $\tilde{E}_{\omega,\varepsilon}$ is much smaller than $E_{\omega,\varepsilon}$ and $\tilde{I}_{\omega,\varepsilon} - I_{\omega,\varepsilon}$ is of the same order of magnitude as $E_{\omega,\varepsilon}$ with all these functions defined in an slightly smaller domain than the original ones.

According to Proposition 2.1, the Hamiltonian of $g_{\omega,\varepsilon}^{-1} \circ f_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}$ is

$$F_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} - G_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} + G_{\omega,\varepsilon} \circ f_{\omega,\varepsilon}^{-1} \circ g_{\omega,\varepsilon}. \tag{3.14}$$

Heuristically, assuming that $G_{\omega,\varepsilon}$ and $E_{\omega,\varepsilon}$ are small and of the same order—and therefore that $g_{\omega,\varepsilon} - \mathrm{Id}$ and $f_{\omega,\varepsilon} - i_{\omega,\varepsilon}$ are small, where $i_{\omega,\varepsilon}$ is the integrable part of $f_{\omega,\varepsilon}$ as in (3.6)—the main terms in (3.14) are

$$F_{\omega,\varepsilon} - G_{\omega,\varepsilon} + G_{\omega,\varepsilon} \circ i_{\omega,\varepsilon}^{-1}.$$

Hence, to make the new error $\tilde{E}_{\omega,\varepsilon}$ zero in this linear approximation, we need to determine $G_{\omega,\varepsilon}$ in such a way that these main terms give just an integrable system (which we will call $\tilde{I}_{\omega,\varepsilon}$). This is formulated as the equation for $G_{\omega,\varepsilon}$, $\tilde{I}_{\omega,\varepsilon}$, given $F_{\omega,\varepsilon}$:

$$\tilde{I}_{\omega,\varepsilon}(p) = F_{\omega,\varepsilon}(p, q) - G_{\omega,\varepsilon}(p, q) + G_{\omega,\varepsilon} \circ i_\omega^{-1}(p, q).$$

Equivalently, we look for an approximate solution of

$$\Delta_{\omega,\varepsilon}(p) = E_{\omega,\varepsilon}(p, q) - G_{\omega,\varepsilon}(p, q) + G_{\omega,\varepsilon} \circ i_\omega^{-1}(p, q), \tag{3.15}$$

where $\Delta_{\omega,\varepsilon}(p) := \tilde{I}_{\omega,\varepsilon}(p) - I_{\omega,\varepsilon}(p)$.

This approximate solution will be used to construct a $g_{\omega,\varepsilon}$, which will lead to a Hamiltonian which is much closer to integrable.

Indeed, the approximate solution of (3.15) will be chosen as an exact solution of

$$\Delta_{\omega,\varepsilon}(p) = E_{\omega,\varepsilon}(p, q) - G_{\omega,\varepsilon}(p, q) + G_{\omega,\varepsilon}(p, q - \omega_0) \tag{3.16}$$

which can be solved by taking Fourier coefficients. We will show that, if we restrict ourselves to a domain $\Sigma_{\Omega,\alpha,\tilde{\beta},\tilde{U}}$, with α very small, the solutions of (3.16) solve (3.15) up to errors that can be controlled by α. Then, the system will be reduced very approximately to a new integrable one. If the frequency function Ω is nondegenerate, we can apply the implicit function theorem and express the domain in terms of the new frequency function $\tilde{\Omega}$. We call attention to the fact that it is only in this last step that the nondegeneracy of the frequency function is used.

To justify the above heuristic argument, we will just find the $g_{\omega,\varepsilon}$ obtained by the procedure detailed above and estimate rigorously the remainder after we conjugate the original problem with it. This task will take most of the present section. We will collect all the estimates systematically and, at the end of the section we will formulate the final result precisely. Once we have these results, we will also need to estimate how the integrable part has changed and, in particular, how much the folding surface Σ and its parameterization Υ introduced in (3.3) have changed. This is the task we will undertake in the next section. Then, in a subsequent section, we will show that the procedure can be iterated indefinitely (when some of the arbitrary choices are made appropriately), and that the transformations converge to a limiting transformation that reduces the system to integrable.

3.4. The iterative step. Estimates. In this subsection, we present detailed quantitative estimates for the iterative step that we described informally in the previous section.

Following standard practice, we denote by K sufficiently large positive constants that depend only on the dimension, the number ω_0, and other elements that remain constant during the proof and denote by K^{-1} all sufficiently small positive constants. We will also need to assume that some quantities related to the integrable part of the system remain bounded under the iteration. We will use K_1, K_2 for these constants that depend on the integrable part. The constants K may depend on these K_1, K_2 but not vice versa. When we discuss the iteration, we will see that these K_1, K_2 are chosen in the first step and then they remain unaltered. In particular, we will need to assume that the constants A and B that quantify the nondegeneracy assumptions (3.12) satisfy

$$A \leq K_1, \quad B \leq K_2. \tag{3.17}$$

Recall that the goal was, given a Hamiltonian with an error term E, defined in a set $\Sigma_{\Omega,\alpha,\beta,U}$ of the form defined in (3.9), to perform a transformation that has an error term $\tilde{E}$ which is much smaller even if defined in a smaller set $\Sigma_{\tilde{\Omega},\tilde{\alpha},\tilde{\beta},\tilde{U}}$.

As it turns out, we will take a number σ and take $\tilde{\beta} = \beta - \alpha - 4\sigma$, $\tilde{U} = U_{4\sigma}$. At the n step σ_n will be $\sigma_0 2^{-n}$, but α will have to decrease superexponentially.

Our goal will be to show that, under appropriate hypotheses, which we will assume inductively, we can perform the transformation and obtain estimates of the form

$$\|\tilde{E}\|_{\tilde{\Omega},\tilde{\alpha},\tilde{\beta},\tilde{U}} \leq K\sigma^{-\tau}\|E\|_{\Omega,\alpha,\beta,U}(\|E\|_{\Omega,\alpha,\beta,U} + \tilde{\alpha}) \tag{3.18}$$

for some fixed positive number τ (we will show later that it suffices to take $\tau = 2\nu + 3$ where $\nu = \theta + d - 1$, and θ is the Diophantine exponent of ω_0).

We will also establish that Υ and $\tilde{\Upsilon}$—the parameterizations (3.3) of the surfaces $\Omega = \omega_0$ and $\tilde{\Omega} = \omega_0$, respectively—are defined in very similar domains and differ by a small amount

$$\|\Upsilon - \tilde{\Upsilon}\|_{\tilde{U}} \leq K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U}. \tag{3.19}$$

The proof will be conveniently divided into two parts. In the first one, we obtain estimates in terms of the old domains parameterized by Ω and α. In this first part—culminated in Lemma 3.6—we will not need to use any nondegeneracy hypothesis in

Ω and indeed ω and ε will just go along for the ride. In a second part of the inductive step, we adjust the domains to the new frequency map. This part will require that we assume that Ω is nondegenerate and we will have to lose some domain in ω. This division is natural since the first part is exactly the same as that used in the proof of Theorem 1.6.

Remark. For the experts in KAM theory, we call attention to the fact that the right-hand side of (3.18) is not quadratic in $\|E\|_{\Omega,\alpha,\beta,U}$—the size of the error. Nevertheless, the linear term is multiplied by the number $\tilde{\alpha}$. As we will see in the following subsection, as $\tilde{\alpha}$ goes to zero superexponentially with the number of steps taken, it is possible to recover the superexponential convergence of KAM theory that beats the small divisors.

As is customary in KAM theory, in order to be able to carry out the iterative step, we will need to assume that certain quantities are sufficiently small with respect to others—so that, for example, compositions have domains that match, implicit function theorems can be applied, etc. As it will turn out all the conditions necessary to perform the iterative step will be implied by smallness conditions of $\|E\|_{\Omega,\alpha,\beta,U}$ with respect to other quantities. Since the iterative step implies that this goes to zero extremely fast, the conditions will be recovered from one step to the next.

Hence, for the proof of Theorem 1.5, the main result of this subsection will be Lemma 3.7 below, which states that, under some explicit conditions, the iterative step can be performed and that the result satisfies (3.18) and (3.19).

Since the proof of Lemma 3.7 will consist in walking through the steps outlined before and just record the conditions needed for them to go through, it is natural to start with the proof of the lemma and postpone its precise statement.

Using Proposition 2.1, the Hamiltonian of $g^{-1}_{\omega,\varepsilon} \circ f_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}$—if it is possible to define all the compositions—is $I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} + E_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} - G_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} + G_{\omega,\varepsilon} \circ f^{-1}_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}$, which adding and subtracting appropriate terms becomes

$$\begin{aligned}
&\overline{I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}} + (I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} - \overline{I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}})\\
&+\overline{E_{\omega,\varepsilon}}\\
&+(E_{\omega,\varepsilon} - \overline{E_{\omega,\varepsilon}}) + (E_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} - E_{\omega,\varepsilon})\\
&-G_{\omega,\varepsilon} + (-G_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} + G_{\omega,\varepsilon})\\
&+G_{\omega,\varepsilon} \circ T^0 + (G_{\omega,\varepsilon} \circ i^{-1}_{\omega,\varepsilon} - G_{\omega,\varepsilon} \circ T^0) + (G_{\omega,\varepsilon} \circ f^{-1}_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} - G_{\omega,\varepsilon} \circ i^{-1}_{\omega,\varepsilon}),
\end{aligned} \tag{3.20}$$

where we have used the notation $\overline{\quad\quad}$ to indicate average over the q variables and $T^0(p, q) = (p, q - \omega_0)$.

The main idea will be to show that it is possible to choose $G_{\omega,\varepsilon}$ in such a way that the first terms in the last three lines of (3.20) add to zero. That is,

$$E_{\omega,\varepsilon} - \overline{E_{\omega,\varepsilon}} - G_{\omega,\varepsilon} + G_{\omega,\varepsilon} \circ T^0 = 0 \tag{3.21}$$

and that this $G_{\omega,\varepsilon}$ satisfies estimates which will guarantee that the compositions we used are indeed defined. (We call attention to the fact that (3.21) is the linearized equation that always appears in KAM theory.) Then, the transformed system will have an integrable part $\tilde{I}_{\omega,\varepsilon} = \overline{I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}} + \overline{E_{\omega,\varepsilon}}$ and the other terms appearing in (3.20) will be the error part of the new Hamiltonian. We will estimate them and show that, in a precise sense, they will be smaller than the other ones.

Remark. For the experts in KAM theory, we note that this procedure has two error terms that are linear in G—and hence first order in E— namely $(G_{\omega,\varepsilon} \circ i^{-1}_{\omega,\varepsilon} - G_{\omega,\varepsilon} \circ T^0)$ and $(I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon} - \overline{I_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}})$—recall that I will not be converging to zero.

Even if full details will be given later, we advance that for the first term, in the domains that we are considering, $i_{\omega,\varepsilon}^{-1}$ and T^0 are indeed close and the distance is measured by $\tilde{\alpha}$. The mean value theorem will give an estimate that contains the factor $\|E\|\tilde{\alpha}$ multiplied by the small divisors. This is the estimate that appears in one of the terms in (3.18). The second term will turn out to be quadratic because of the fact that $g_{\omega,\varepsilon}$ is exact symplectic. This is the only place in all the estimates where we use that the maps are exact symplectic.

As usual in KAM theory, we start by obtaining bounds on $G_{\omega,\varepsilon}$ and we will use them to obtain bounds on all the other terms.

Lemma 3.2. *For any $E_{\omega,\varepsilon}(p,q)$ defined in $\Sigma_{\Omega,\alpha,\beta,U}$, we can find unique $\Delta_{\omega,\varepsilon}(p)$, $G_{\omega,\varepsilon}(p,q)$ satisfying*

$$\Delta_{\omega,\varepsilon}(p) = E_{\omega,\varepsilon}(p,q) - G_{\omega,\varepsilon}(p,q) + G_{\omega,\varepsilon}(p,q-\omega_0)$$
$$\int_{\mathbb{T}^d} G_{\omega,\varepsilon}(p,q)\,dq = 0.$$

Moreover, these Δ, G satisfy

$$\|G\|_{\beta-\sigma,U} \le K\sigma^{-\nu}\,\|E\|_{\beta,U}\,, \qquad \|\Delta\|_{\beta,U} \le \|E\|_{\beta,U}\,, \tag{3.22}$$

where $\nu = \theta + d - 1$.

Proof. The proof is quite standard. We note that integrating in q we have

$$\Delta_{\omega,\varepsilon}(p) = \overline{E_{\omega,\varepsilon}}(p) := \int_{\mathbb{T}^d} dq\, E_{\omega,\varepsilon}(p,q), \tag{3.23}$$

hence, the first estimate in (3.22) follows.

If we take Fourier transforms in the variable q we obtain

$$\hat{G}_{\omega,\varepsilon;k}(p) = \frac{1}{(e^{-2\pi i k\cdot\omega_0} - 1)}\hat{E}_{\omega,\varepsilon;k}(p). \tag{3.24}$$

By the Cauchy estimates of Lemma 3.1, we have $|\hat{E}_{\omega,\varepsilon;k}(p)| \le Ke^{-2\pi\beta|k|}\,\|E\|_{\beta,U}$ and, by the Diophantine assumptions, $|e^{-2\pi i k\omega_0} - 1|^{-1} \le C|k|^{\theta-1}$. Hence,

$$|\hat{G}_{\omega,\varepsilon;k}(p)| \le K|k|^{\theta-1}e^{-2\pi\beta|k|}\,\|E\|_{\beta,U}$$

and, therefore

$$\begin{aligned}\|G\|_{\beta-\sigma,U} &\le \textstyle\sum_{k\in\mathbb{Z}^d} |\hat{G}_{\omega,\varepsilon;k}(p)|e^{2\pi(\beta-\sigma)|k|} \le K\left(\sum_{k\in\mathbb{Z}^d}|k|^{\theta-1}e^{-2\pi\sigma|k|}\right)\|E\|_{\beta,U}\\ &\le K\left(\textstyle\sum_{l\in\natural}|l|^{\theta-1+d-1}e^{-2\pi\sigma l}\right)\|E\|_{\beta,U} \le K\sigma^{-\nu}\,\|E\|_{\beta,U}\,,\end{aligned}$$

where $\nu = \theta + d - 1$.

We refer to [SM] for more details but point out that it is possible to obtain better exponents in σ (see, e.g., [Ru]). Of course, since the rest of the proof goes through for any exponent, this does not affect the subsequent reasoning. □

A small generalization of these estimates is in the following proposition.

Proposition 3.3. *With the notation of Lemma* 3.2

$$\|\nabla^m G\|_{\beta-\sigma,U_\sigma} \le K\sigma^{-\nu-m}\,\|E\|_{\beta,U}\,. \tag{3.25}$$

Proof. Using Lemma 3.1 and (3.24) we obtain that, for $(\omega, \varepsilon, p) \in U_\sigma$, we have

$$|\partial_p^i \hat{G}_{\omega,\varepsilon;k}(p)| \le K\sigma^{-(i+\theta-1)} e^{-2\pi|k|\beta} \|E\|_{\beta,U}. \tag{3.26}$$

Similarly, we have

$$|\partial_q^j(\hat{G}_{\omega,\varepsilon;k}(p)e^{2\pi i k\cdot q})| \le K|k|^j |\hat{G}_{\omega,\varepsilon;k}(p)| \le K|k|^{j+\theta-1} e^{-2\pi|k|\beta} \|E\|_{\beta,U}. \tag{3.27}$$

On $\Sigma_{\beta-\sigma,U_\sigma}$ we have $|\Im q| \le \beta - \sigma$ and hence $|e^{2\pi i k\cdot q}| \le e^{2\pi|k|(\beta-\sigma)}$. Therefore, using the above estimates (3.26) and (3.27) in the same way as in Lemma 3.2, we obtain the desired result. □

Now, we can prove estimates for the flow of $G_{\omega,\varepsilon}$.

PROPOSITION 3.4. *Assume that the conditions of Proposition* 3.3 *are met and that, furthermore,*

$$K\sigma^{-\nu-1} \|E\|_{\beta,U} \le \sigma/2. \tag{3.28}$$

Then

(i) *for* $(\omega, \varepsilon, p, q) \in \Sigma_{\beta-2\sigma,U_{2\sigma}}$, *the flow* $g_{\omega,\varepsilon}(p,q)$ *generated by the Hamiltonian* $G_{\omega,\varepsilon}$ *is well defined, and* $(\omega, \varepsilon, g_{\omega,\varepsilon}(p,q)) \in \Sigma_{\beta-\sigma,U_\sigma}$;

(ii) $\|g - \mathrm{Id}\|_{\beta-2\sigma,U_{2\sigma}} \le \|\nabla G\|_{\beta-\sigma,U_\sigma} \le K\sigma^{-\nu-1} \|E\|_{\beta,U}$.

Proof. It follows from hypothesis (3.28), Proposition 3.3, and the local existence theorem for solutions of ODEs. □

From now on, we will assume that (3.28) holds, and we will proceed to estimate the terms in (3.20).

By Proposition 3.4, the compositions $G_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}$, $E_{\omega,\varepsilon} \circ g_{\omega,\varepsilon}$ are well defined on $\Sigma_{\beta-2\sigma,U_{2\sigma}}$. Using the mean value theorem and Cauchy inequalities from Lemma 3.1, we can bound

$$\|G - G\circ g\|_{\beta-2\sigma,U_{2\sigma}} \le \|\nabla G\|_{\beta-\sigma,U_\sigma} \|g - \mathrm{Id}\|_{\beta-2\sigma,U_{2\sigma}} \le K\sigma^{-2\nu-2} \|E\|^2_{\beta,U}, \tag{3.29}$$

$$\|E - E\circ g\|_{\beta-2\sigma,U_{2\sigma}} \le \|\nabla E\|_{\beta-\sigma,U_\sigma} \|g - \mathrm{Id}\|_{\beta-2\sigma,U_{2\sigma}} \le K\sigma^{-\nu-2} \|E\|^2_{\beta,U}. \tag{3.30}$$

These estimates show that two of the terms in (3.20) are quadratically small in the original error.

Now, we turn to estimate the last term in (3.20), which, as we will show, will also be quadratic in $\|E\|$. The reason is that $f_{\omega,\varepsilon}$ and $i_{\omega,\varepsilon}$ satisfy differential equations whose difference can be controlled by $\|E\|$ and the same initial conditions. Hence, $\|f^{-1} - i^{-1}\| \le K\|E\|$ under some mild extra assumptions that guarantee that domains match, etc., and we can now apply the mean value theorem. The precise details are a walk through the standard proof of the existence and uniqueness for ODEs, as we detail below.

First, we recall that $i_{\omega,\varepsilon}$ has the form (3.6): $i_{\omega,\varepsilon}(p,q) = (p, q + \Omega(\omega,\varepsilon,p))$, with $\Omega(\omega,\varepsilon,p)$ given in (3.11), and we note that $i^{-1}_{\omega,\varepsilon}(p,q) = (p, q - \Omega(\omega,\varepsilon,p))$. Hence, for

$$\|i - T_0\|_U = \left\|i^{-1} - T_0\right\|_U = \|\Omega - \omega_0\|_U \le \alpha \tag{3.31}$$

we have

$$(\omega,\varepsilon,p,q) \in \Sigma_{\beta-\alpha,U} \Longrightarrow (\omega,\varepsilon,i_{\omega,\varepsilon}(p,q)), \left(\omega,\varepsilon,i^{-1}_{\omega,\varepsilon}(p,q)\right) \in \Sigma_{\beta,U}. \tag{3.32}$$

Assuming

$$\left\|\frac{\partial\Omega}{\partial p}\right\|_U \le K_3 \tag{3.33}$$

(where without loss of generality, we assume, to simplify some formulas that $K_3 > 1$), we can bound

$$\|\nabla i\|_U = \left\|\nabla i^{-1}\right\|_U \le K. \tag{3.34}$$

We recall now that $f_{\omega,\varepsilon}$ is the solution of

$$\begin{aligned} f_{\omega,\varepsilon}(x) &= f_{\omega,0}(x) + \textstyle\int_0^\varepsilon ds\, \mathcal{F}_{\omega,s} \circ f_{\omega,s}(x) \\ &= f_{\omega,0}(x) + \textstyle\int_0^\varepsilon ds\, \left[\mathcal{I}_{\omega,\varepsilon} \circ f_{\omega,s}(x) + \mathcal{E}_{\omega,\varepsilon} \circ f_{\omega,s}(x)\right] \end{aligned} \tag{3.35}$$

while $i_{\omega,\varepsilon}$ satisfies $i_{\omega,\varepsilon}(x) = i_{\omega,0}(x) + \int_0^\varepsilon ds\, \mathcal{I}_{\omega,s} \circ i_{\omega,s}(x)$, with $f_{\omega,0}(x) = i_{\omega,0}(x)$. By hypothesis (3.28), using standard arguments of ODEs based on the Gronwall inequality, we get that for $(\omega,\varepsilon,p,q) \in \Sigma_{\beta-\alpha-2\sigma,U_{2\sigma}}$, the flow $f_{\omega,\varepsilon}(p,q)$ is well defined, and satisfies

$$(\omega,\varepsilon,p,q) \in \Sigma_{\beta-\alpha-2\sigma,U_{2\sigma}} \Longrightarrow (\omega,\varepsilon,f_{\omega,\varepsilon}(p,q)) \in \Sigma_{\beta-\sigma,U_\sigma}, \tag{3.36}$$

$$\|f - i\|_{\beta-\alpha-2\sigma,U_{2\sigma}} \le e^{K_3} \|\nabla E\|_{\beta-\sigma,U_\sigma} \le K\sigma^{-1} \|E\|_{\beta,U}. \tag{3.37}$$

From (3.34), and Lemma 3.1 applied to (3.37), we can bound $\nabla f_{\omega,\varepsilon}$:

$$\|\nabla f\|_{\beta-\alpha-3\sigma,U_{3\sigma}} \le \|\nabla i\|_U + \|\nabla(f-i)\|_{\beta-\alpha-3\sigma,U_{3\sigma}} \le K. \tag{3.38}$$

Applying the implicit function theorem to the estimates above, it turns out that for $(\omega,\varepsilon,p,q) \in \Sigma_{\beta-\alpha-2\sigma,U_{2\sigma}}$, $f^{-1}_{\omega,\varepsilon}(p,q)$ is well defined, satisfies $\left(\omega,\varepsilon,f^{-1}_{\omega,\varepsilon}(p,q)\right) \in \Sigma_{\beta-\sigma,U_\sigma}$, and

$$\left\|f^{-1} - i^{-1}\right\|_{\beta-\alpha-2\sigma,U_{2\sigma}} \le K\sigma^{-1}\| \, \|E\|_{\beta,U}. \tag{3.39}$$

As before, from (3.34), and Lemma 3.1 applied to (3.39), we can bound $\nabla f^{-1}_{\omega,\varepsilon}$:

$$\left\|\nabla f^{-1}\right\|_{\beta-\alpha-3\sigma,U_{3\sigma}} \le K. \tag{3.40}$$

Using the mean value theorem, (3.40), and the bounds on $g_{\omega,\varepsilon} - \mathrm{Id}$ established in Proposition 3.4, we obtain

$$\left\|f^{-1} - f^{-1} \circ g\right\|_{\beta-\alpha-3\sigma,U_{3\sigma}} \le K\sigma^{-\nu-1} \|E\|_{\beta,U}. \tag{3.41}$$

Putting together (3.39) and (3.41), by the triangle inequality, we obtain

$$\left\|f^{-1} \circ g - i^{-1}\right\|_{\beta-\alpha-3\sigma,U_{3\sigma}} \le K\sigma^{-\nu-1} \|E\|_{\beta,U}. \tag{3.42}$$

Using the mean value theorem, the estimates in Proposition 3.3, and (3.42), we can bound the last term in (3.20) as

$$\left\|G \circ f^{-1} \circ g - G \circ i^{-1}\right\|_{\beta-\alpha-3\sigma,U_{3\sigma}} \le K\sigma^{-2\nu-2} \|E\|^2_{\beta,U}. \tag{3.43}$$

Now we turn our attention to the first term in (3.20). It will depend on the approximate expression $g^0_{\omega,\varepsilon} = \mathrm{Id} + \int_0^\varepsilon ds\, \mathcal{G}_{\omega,s}$ for $g_{\omega,\varepsilon}$:

$$g^0_{\omega,\varepsilon}(p,q) = \left(p - \int_0^\varepsilon ds\, \frac{\partial}{\partial q} G_{\omega,s}(p,q)\ , q + \int_0^\varepsilon ds\, \frac{\partial}{\partial p} G_{\omega,s}(p,q)\right). \tag{3.44}$$

Proposition 3.5. *Under our standing hypotheses, we have*

$$\left\|g - g^0\right\|_{\beta-2\sigma,U_{2\sigma}} \leq K\sigma^{-2\nu-3}\left\|E\right\|^2_{\beta,U}.$$

Proof. Note that our standing assumptions imply

$$\left\|g^0 - \mathrm{Id}\right\|_{\beta-\sigma,U_\sigma} \leq \left\|\nabla G\right\|_{\beta-\sigma,U_\sigma} \leq K\sigma^{-\nu-1}\left\|E\right\|_{\beta,U}$$

and consequently $\left(\omega, \varepsilon, g^0_{\omega,\varepsilon}(p,q)\right) \in \Sigma_{\beta-\sigma,U_\sigma}$ for $(\omega,\varepsilon,p,q) \in \Sigma_{\beta-\alpha-2\sigma,U_{2\sigma}}$.

We can write $g_{\omega,\varepsilon}$ as the solution of a fixed point problem. Namely,

$$g_{\omega,\varepsilon} = \mathrm{Id} + \int_0^\varepsilon ds\, \mathcal{G}_{\omega,s} \circ g_{\omega,s} \equiv \mathcal{T}(g)_{\omega,\varepsilon},$$

and we have the identity

$$\mathcal{T}(g^0)_{\omega,\varepsilon} - g^0_{\omega,\varepsilon} = \int_0^\varepsilon ds\, [\mathcal{G}_{\omega,s} \circ g^0_{\omega,s} - \mathcal{G}_{\omega,s}].$$

If we estimate the integrand of the right-hand side (R.H.S.) by the mean value theorem, we have

$$\begin{aligned}\left\|\mathcal{T}(g^0) - g^0\right\|_{\beta-2\sigma,U_{2\sigma}} &\leq \left\|\nabla^2 G\right\|_{\beta-\sigma,U_\sigma}\left\|g^0 - \mathrm{Id}\right\|_{\beta-\sigma,U_\sigma} \\ &\leq K\sigma^{-2\nu-3}\left\|E\right\|^2_{\beta,U}.\end{aligned} \tag{3.45}$$

We also obtain, under (3.28), that $\mathcal{T}$ is a contraction of factor 1/2. Hence, there is a fixed point of $\mathcal{T}$ whose distance from $g^0_{\omega,\varepsilon}$ is not bigger than $1/(1-1/2) = 2$ times the R.H.S. of (3.45). □

We note that, because $I_{\omega,\varepsilon}(x)$ does not depend on q, denoting by Π_p, Π_q the projections on the p and q components, respectively, we have for $x = (p,q)$

$$\begin{aligned}I_{\omega,\varepsilon}(g_{\omega,\varepsilon}(x)) &= I_{\omega,\varepsilon}(\Pi_p g_{\omega,\varepsilon}(x)) \\ &= I_{\omega,\varepsilon}(p) + \partial_p I_{\omega,\varepsilon}(p)\Pi_p\left[g_{\omega,\varepsilon}(x) - x\right] + R_2\left(\omega,\varepsilon,x,g_{\omega,\varepsilon}(x)\right) \\ &= I_{\omega,\varepsilon}(p) + \partial_p I_{\omega,\varepsilon}(p)\Pi_p\left[g^0_{\omega,\varepsilon}(x) - x\right] \\ &\quad + \partial_p I_{\omega,\varepsilon}(p)\Pi_p\left[g_{\omega,\varepsilon}(x) - g^0_{\omega,\varepsilon}(x)\right] + R_2\left(\omega,\varepsilon,x,g_{\omega,\varepsilon}(x)\right),\end{aligned} \tag{3.46}$$

where we have denoted by R_2 the remainder of the second order Taylor expansion in p.

Note that $\Pi_p\left[g^0_{\omega,\varepsilon}(x) - x\right] = \partial_q G_{\omega,\varepsilon}(x)$ (see (3.44)) and that $\overline{\partial_q G_{\omega,\varepsilon}} = 0$ since q is a periodic variable. Hence, observing that $\partial_p I$ is independent of q, we obtain

$$\overline{\partial_p I \int_0^\varepsilon ds\, \partial_q G_{\omega,s}} = 0. \tag{3.47}$$

That is, the second term in the R.H.S. of the formula of (3.46) has zero average. We call attention to the fact that this is the only part in the whole proof of the estimates where we use the exact symplectic character of the deformation, which is equivalent to the fact that G is a function on the annulus and not just on the universal cover.

Since $I_{\omega,\varepsilon}$ depends only on p we have that $\overline{I_{\omega,\varepsilon}} = I_{\omega,\varepsilon}$.

Under the assumption

$$\left\|\nabla^2 I\right\|_{\beta-\sigma,U_\sigma} \le K_4 \tag{3.48}$$

we can bound the last two terms in (3.46) by terms that are quadratic in $\|E\|$.

Since the last two terms in (3.46) are the only ones that contribute to $I_{\omega,\varepsilon}\circ g_{\omega,\varepsilon} - \overline{I_{\omega,\varepsilon}\circ g_{\omega,\varepsilon}}$, we obtain from Proposition 3.5

$$\left\|I\circ g - \overline{I\circ g}\right\|_{\beta-2\sigma,U_{2\sigma}} \le K\sigma^{-2\nu-3}\left\|E\right\|^2_{\beta,U}. \tag{3.49}$$

The only term in (3.20) that remains to be estimated is $G_{\omega,\varepsilon}\circ i^{-1}_{\omega,\varepsilon} - G_{\omega,\varepsilon}\circ T^0$. We note that, by (3.31), we have

$$\left\|i^{-1} - T^0\right\|_U \le \alpha.$$

Therefore, using the estimates in Proposition 3.3,

$$\left\|G\circ i^{-1} - G\circ T^0\right\|_{\sigma,U_{\beta-\alpha-\sigma}} \le K\sigma^{-\nu-1}\alpha\left\|E\right\|_{\beta,U}. \tag{3.50}$$

If we add the estimates in (3.29), (3.30), (3.43), (3.49), and (3.50), for the terms that have to be bounded in (3.20), and claim them only in the domain $\Sigma_{\beta-\alpha-4\sigma,U_{4\sigma}}$, which is smaller than any of the domains in which we have bounds, we obtain

$$\left\|\tilde{E}\right\|_{\beta-\alpha-4\sigma,U_{4\sigma}} \le K\sigma^{-\tau}\left\|E\right\|_{\beta,U}\left(\left\|E\right\|_{\beta,U} + \alpha\right), \tag{3.51}$$

where $\tau := 2\nu+3$ and $\left\|\Omega-\omega_0\right\|_U \le \alpha$.

We also notice that from Proposition 3.4 and (3.36), it follows that if $(\omega,\varepsilon,p,q)\in \Sigma_{\beta-\alpha-4\sigma,U_{4\sigma}}$, then $\left(\omega,\varepsilon,g^{-1}_{\omega,\varepsilon}\circ f_{\omega,\varepsilon}\circ g_{\omega,\varepsilon}(p,q)\right)\in\Sigma_{\beta-\sigma,U_\sigma}$.

On the set $\Sigma_{\Omega,\alpha,\beta\alpha-4\sigma,U_{4\sigma}}$ introduced in (3.9), (3.51) reads as

$$\|\tilde{E}\|_{\Omega,\alpha,\beta-\alpha-4\sigma,U_{4\sigma}} \le K\sigma^{-\tau}\|E\|_{\Omega,\alpha,\beta,U}(\|E\|_{\Omega,\alpha,\beta,U}+\alpha). \tag{3.52}$$

This is very similar to the estimates desired in (3.18) and it only differs from them in the fact that the norm in the left-hand side (L.H.S.) of (3.52) is referred to as the domain specified by Ω and not by $\tilde{\Omega}$.

To remedy that, we will estimate the change in Ω and the attendant change in the parameterizations Υ of the surface and the domain Σ. Using that the frequency function Ω is nondegenerate, this will allow us to transform the expression of the domain in which we have improved estimates into an expression involving the new frequency function.

We will find it convenient to state formally what we have already accomplished without using nondegeneracy conditions in Ω. We call attention that this lemma will also play an important role in the proof of Theorem 1.6. Later, we will prove Lemma 3.7 that takes into account the change in the frequency function and which indeed uses the nondegeneracy assumptions in Ω.

LEMMA 3.6. *Given the Hamiltonian $F = I + E$ of $f_{\omega,\varepsilon}$ introduced in* (3.10), *choose G, Δ as given by Lemma* 3.2, *and consider the new Hamiltonian $\tilde{F} = \tilde{I} + \tilde{E}$ of $g^{-1}_{\omega,\varepsilon}\circ f_{\omega,\varepsilon}\circ g_{\omega,\varepsilon}$ as given in* (3.13). *Assume that σ is such that* (3.28), (3.33), *and* (3.48) *are met, and let $\tau = 2\nu+3$. Then*

$$\|\tilde{E}\|_{\Omega,\alpha,\beta-\alpha-4\sigma,U_{4\sigma}} \le K\sigma^{-\tau}\|E\|_{\Omega,\alpha,\beta,U}(\|E\|_{\Omega,\alpha,\beta,U}+\alpha), \tag{3.53}$$

$$\|\Delta\|_{\Omega,\alpha,\beta,U} \le \|E\|_{\Omega,\alpha,\beta,U}, \qquad \|\nabla\Delta\|_{\Omega,\alpha,\beta-\sigma,U_\sigma} \le K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U}. \tag{3.54}$$

The way of interpreting these estimates is that (3.53) indicates that, after the transformation, the resulting Hamiltonian is essentially an integrable one (albeit in a smaller domain): the right-hand side of (3.53) consists on two terms, one of which is quadratic in $\|E\|$ and the other one contains $\|E\|\alpha$. If we choose α sufficiently small, we will be able to make the R.H.S. of (3.53) much smaller than the original one. This will overcome the small divisors $\sigma^{-\tau}$.

We call attention to the fact that Lemma 3.6 does not need the nondegeneracy assumption on Ω and that it does not lose any domain in the parameters. This lemma will a basic tool for the estimates of the inductive steps both in the proof of the KAM theorem and in the justification of Greene's criterion. The difference between the two results will be that the inductive steps will have different domain losses and that we will have to apply them repeatedly in different ways, losing domain at different rates.

3.5. The KAM inductive step. Geometry of domains. To complete the work for the bounds of the inductive step in the KAM theorem, we need to study the change in Ω, the surface Σ defined by $\Omega = \omega_0$ and its natural parameterization Υ defined in (3.3). In particular, we will need to provide estimates for the changes of the bounds in (3.12) that quantify the nondegeneracy assumptions. Since we are also taking into account the derivative of Ω with respect to ω, instead of (3.33), we are going to assume

$$\left\|\frac{\partial\Omega}{\partial p}\right\|_U \leq K_3, \qquad \left\|\frac{\partial\Omega}{\partial\omega}\right\|_U \leq K_3. \tag{3.55}$$

Again, we emphasize that most of the results in this section are true for arbitrary d. The only exception is (iv) in Lemma 3.7 below.

Given the estimates that we have on Δ, it will be very easy to estimate the change in Ω and all the other estimates will follow by an application of the implicit function theorem. We note that since Δ is small, and Ω depends linearly on the integrable part, the change in Ω will be of the same order of magnitude and hence also small. All the changes in the surface and in the parameterization will be small and hence can be estimated by $\|E\|$ possibly multiplied by some factors that come from the fact that we have to involve derivatives and control them by Cauchy estimates.

More precisely, we have the following lemma.

Lemma 3.7. *Let Ω be the frequency function* (3.11) *for the family $f_{\omega,\varepsilon}$* (3.10) *defined on $\Sigma_{\Omega,\alpha,\beta,U}$ as in* (3.9). *Let Δ be given by* (3.23) *and let σ be a positive number. Assume that* (3.17), (3.28), (3.55), *and* (3.48) *hold. Consider $\tilde{\Omega}$, the new frequency function defined by*

$$\tilde{\Omega}(\omega,\varepsilon,p) = \Omega(\omega,\varepsilon,p) + \int_0^\varepsilon ds\, \frac{\partial}{\partial p}\Delta(\omega,s,p). \tag{3.56}$$

Denote by Υ and $\tilde{\Upsilon}$ the parameterizations (3.3) *corresponding to Ω and $\tilde{\Omega}$.*

Then, for any $\tilde{\alpha} \leq \alpha$ satisfying

$$K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U} \leq \tilde{\alpha} \tag{3.57}$$

we have

(i) $\|\Omega - \tilde{\Omega}\|_{U_\sigma} \leq K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U} \leq \tilde{\alpha}$;

(ii) *for $\tilde{\alpha}$ as before, $\tilde{\beta} = \beta - \alpha - 4\sigma$, $\tilde{U} \equiv U_{4\sigma}$, we have*

$$\Sigma_{\tilde{\Omega},\tilde{\alpha},\tilde{\beta},\tilde{U}} \subset \Sigma_{\Omega,2\alpha,\beta-4\sigma,U_{4\sigma}};$$

(iii)

$$\left\|\left(\frac{\partial}{\partial\omega}\tilde{\Omega}\right)^{-1}\right\|_{\tilde{U}} \le \left\|\left(\frac{\partial}{\partial\omega}\tilde{\Omega}\right)^{-1}\right\|_{U_{4\sigma}} \le \left\|\left(\frac{\partial}{\partial\omega}\Omega\right)^{-1}\right\|_{U} + K\sigma^{-2}\|E\|_{\Omega,\alpha,\beta,U};$$

(iv) *when* $d = 1$,

$$\left\|\left(\frac{\partial^2}{\partial p^2}\tilde{\Omega}\right)^{-1}\right\|_{\tilde{U}} \le \left\|\left(\frac{\partial^2}{\partial p^2}\tilde{\Omega}\right)^{-1}\right\|_{U_{4\sigma}} \le \left\|\left(\frac{\partial^2}{\partial p^2}\Omega\right)^{-1}\right\|_{U} + K\sigma^{-3}\|E\|_{\Omega,\alpha,\beta,U};$$

(v) $\|\Upsilon - \tilde{\Upsilon}\|_{\tilde{U}} \le K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U}$;

(vi) *the inequalities* (3.18) *hold, that is, for* $\tau = 2\nu + 3$

$$\|\tilde{E}\|_{\tilde{\Omega},\tilde{\alpha},\tilde{\beta},\tilde{U}} \le K\sigma^{-\tau}\|E\|_{\Omega,\alpha,\beta,U}(\|E\|_{\Omega,\alpha,\beta,U} + \tilde{\alpha}).$$

Proof. Part (i) follows immediately from the formula (3.56) for $\tilde{\Omega}$ and the estimates that we have for Δ in Lemma 3.2. The last inequality in (i) is just a restatement of (3.28), which is one of the hypotheses of the lemma.

Part (ii) follows because of (3.57).

Parts (iii) and (iv) follow because we can use Cauchy estimates to estimate the derivatives of Δ. Then, we can use Cauchy estimates to bound the derivatives of Ω.

The existence of $\tilde{\Upsilon}$ and its estimates are a very simple consequence of the implicit function theorem. Recall the well-known result that if an analytic function Φ satisfies $|\Phi(0)| \le \varepsilon$ and $|\Phi'|^{-1} \le a$ on a ball around zero of radius $a\varepsilon$ there is one and only one zero in this ball. Moreover, if Φ depends analytically on parameters, the zero depends analytically on parameters. We can apply this result to $\Phi(s) = \Omega(s+\Upsilon(\varepsilon,p),\varepsilon,p)-\omega_0$ and then, the result follows.

Part (vi) is a consequence of the estimates in Lemma 3.6 and part (ii) of this lemma. □

Notice that the only places where we had to consider derivatives with respect to ω are (iii) and (v). Hence, this will be easy to adapt to the situation in the justification of the Greene's criterion where there is some degeneracy in the frequency function.

Remark. Notice also that it is only in these nondegeneracy assumptions that we have to consider the one-dimensional properties of the map. It seems that with some appropriate notion of critical circle in higher dimensions (one has to consider invariant tori with "degenerate torsion"), one could develop an analogous converging KAM process, and a subsequent geometrical interpretation could provide the structure of invariant objects nearby the critical torus.

3.6. Iteration of the KAM inductive step. Convergence. In this subsection, we verify that if we start with a sufficiently small perturbation E, the iterative step can be repeated infinitely many times and, moreover, converges to a solution. The estimates are very similar to those in the paper [Ru2] on the translated curve method. Along the rest of this section, we will assume that $d = 1$.

The main idea is that the loss of domain has to be fast—say, exponentially fast—in the variables q so that we have some domain left. On the other hand, we have to decrease superexponentially fast the variable α which controls the thickness of the approximations to the surface Σ. This will achieve that the $\|E\|$ decreases superexponentially and that, as a consequence, the process can be iterated indefinitely.

We will choose α_n, σ_n, and show that if $\|E^0\|_{\Omega^0,\alpha_0,\beta_0,U^0}$ is small enough, the iterative step described in the previous section can be repeated indefinitely and the transformations converge to a solution that indeed solves the problem.

We point out that these smallness conditions can always be adjusted by switching to another variable $\varepsilon' = \varepsilon\lambda$. If we choose λ small enough, the remainder is made arbitrarily small while all the other parameters in the problem are left unaltered. (That is, when we have families, we can obtain the smallness conditions by considering ε restricted to a small domain.) Of course, when our families are obtained by interpolating between two diffeomorphisms, as in Lemma 2.2, the smallness assumptions in the family can be accomplished by assuming that the diffeomorphisms we are interpolating are close.

We will start by picking $\sigma_n = \sigma^* 2^{-n}$, where we pick $\sigma^* < \beta_0/8$ so that β_n defined in Lemma 3.7 by $\beta_{n+1} = \beta_n - 4\sigma_n$ is bounded away from zero, and $\sigma^* < \delta/8$ so that all the domains $U^{n+1} = U^n_{\sigma_n}$ contain the open domain $U^0_{2\sigma^*}$. Now we will show that it is possible to choose α_n in such a way that if $\|E^0\|_{\Omega^0,\alpha_0,\beta_0,U^0}$ is small enough, the process can be iterated indefinitely and it converges.

Introducing the notation $e_n = \|E_n\|_{\Omega^n,\alpha_n,\beta_n,U^n}$, $a_n = \alpha_{n+1}$, $A = 2^\tau$, $C = K/\sigma^*$, the recursion equation in (vi) of Lemma 3.7 becomes

$$e_{n+1} \le CA^n\, e_n(e_n + a_n). \tag{3.58}$$

We claim the following lemma.

Lemma 3.8. *If e_0 is small enough, it is possible to choose $0 < \rho < 1$ in such a way that setting $a_n = \rho^{2^n}(AB)^{-n}$, for $B > 1$, the conditions for Lemma 3.7 are satisfied for all n and*

$$e_n \le \frac{a_n}{C\,2^n} = \frac{\rho^{2^n}}{C(2AB)^n}. \tag{3.59}$$

Proof. Assume that (3.59) holds for a certain n and that we have chosen a_n as indicated and that the iterative step can be applied at this step.

Then, by (3.58) we have

$$\begin{aligned} e_{n+1} &\le \tfrac{\rho^{2^n}}{C(2AB)^n}\left(\tfrac{\rho^{2^n}}{C(2AB)^n} + \tfrac{\rho^{2^n}}{(AB)^n}\right) CA^n \\ &= \tfrac{\rho^{2^{n+1}}}{C(2AB)^{n+1}}\tfrac{2ABC}{B^n}\left(\tfrac{1}{C\,2^n} + 1\right) \le \tfrac{\rho^{2^{n+1}}}{C(2AB)^{n+1}}\tfrac{4ABC}{B^n}. \end{aligned} \tag{3.60}$$

If $n > N_0(A,B,C)$ we have that

$$\frac{4ABC}{B^n} \le 1 \tag{3.61}$$

so that indeed the formula (3.59) holds for $n+1$.

We also observe that, if a_n and e_n are of the form that we claimed, there is an $N_1(A,B,C) \ge N_0$ so that all the hypotheses (3.17), (3.28), (3.55), (3.48), (3.57) are satisfied for $n > N_1$.

Therefore, it suffices to ensure that e_0 is so small that the iterative step can be performed N_1 times and that the inequalities (3.59) hold for $n \le N_1$. Then, the argument in (3.60) will show that (3.59) continue to hold, and that the hypotheses needed to perform the iterative step and (3.61) hold. □

Clearly, from (3.59), we obtain that the error of the solution goes to zero on the surfaces. Similarly, using the estimates in Lemma 3.7 we can show that the parameterizations Υ of the surface converge. (It suffices to check that the increments are summable.)

Moreover, defining $h^n_{\omega,\varepsilon} = g^0_{\omega,\varepsilon} \circ \cdots \circ g^n_{\omega,\varepsilon}$ we have that

$$
\begin{aligned}
\|h^n - h^{n-1}\|_{\Omega^n,\alpha_n,\beta_n,U^n} &= \|h^{n-1} \circ g^n - h^{n-1}\|_{\Omega^n,\alpha_n,\beta_n,U^n} \\
&\le \sigma_{n-1}^{-1} K \|h^{n-1}\|_{\Omega^{n-1},\alpha_{n-1},\beta_{n-1},U^{n-1}} \|g^n - \operatorname{Id}\|_{\Omega^n,\alpha_n,\beta_n,U^n}.
\end{aligned}
\tag{3.62}
$$

From (3.62) and the estimates in (ii) of Proposition 3.4, it is immediate to show by induction that $\|h^n\|_{\Omega^n,\alpha_n,\beta_n,U^n}$ remains bounded independently of n. Then, using (ii) of Proposition 3.4, the R.H.S. of (3.62) is summable in n. Hence $h^n_{\omega,\varepsilon}$ converges in the limiting domain $\Sigma_{\Omega^\infty,\alpha_\infty,\beta_\infty,U^\infty}$, with $\alpha_\infty = 0$, consisting on the points (ω,ε,p,q) with $(\omega,\varepsilon,p) \in U^\infty = U^0_{2\sigma^*}$ such that $\Omega^\infty(\omega,\varepsilon,p) = \omega_0$ and $|\Im q| \le \beta_\infty = \beta_0 - 2\sigma^* \ge \beta_0/2$.

This finishes the proof of Theorem 1.5.

4. Partial justification of Greene's criterion. To assess numerically the existence of invariant circles, the most frequently used method is the so-called Greene's criterion, formulated in [Gr] for two-dimensional maps.

This criterion asserts that a smooth invariant circle with motion smoothly conjugate to a rotation ω exists if and only if it is possible to find a sequence of periodic orbits of type m/n whose "residue" (that is, the trace of the derivative of the return map minus 2) converges to zero as the m/n converges to ω_0.

As it turns out, this criterion has not been proved to hold; nevertheless, parts of it can be established rigorously.

For standard KAM tori, Mather (see [McK, Section 1.3.2.4]) suggested a method to prove that if KAM tori existed, the residue should go to zero faster than any power of $|\omega - p_n/q_n|$. This method was implemented in [FL], [McK2] for two-dimensional maps to show that the residue is smaller than $\exp(-c|\omega - p_n/q_n|^{-\alpha})$ for some $\alpha > 0$.

The main goal of this section is to prove one of the implications of Greene's criterion for critical circles. We will prove that if a critical circle exists, then any sequence of periodic orbits converging to it has residual converging to zero. We will also show that, if a critical circle exists, indeed there is at least one such sequence. Actually, for any m/n such that $m/n < \omega$, $|m/n - \omega| \ll 1$, we can find at least two periodic orbits of type m/n and, under mild nondegeneracy conditions, at least four.

Again, we will assume in this section that $d = 1$. We note that for higher dimensional maps, in [T1] and [T2] there are versions of Greene's criterion for higher dimensional twist maps (a rigorous justification of one of the implications and numerical evidence, respectively). There are some differences between the proofs in higher dimensional cases and the case considered here of $d = 1$ and we will comment on them after the proof of our results.

The main part of the proof will consist in showing that, in a neighborhood of the invariant circle, it is possible to find changes of variables that reduce the system almost to integrable. Once we have that, the result will follow word for word the result in [FL].

Of course, the estimates near the invariant torus are a more general result than that of the Greene's criterion and they allow us to control not only the behavior of the periodic orbits but also other dynamical objects. Other papers in which similar estimates are obtained for nondegenerate circles are [OS], [PW], [JV], [DG2].

Most of the work has already been done in section 3. The estimates that we will use are the same as those of the iterative step and the only difference is that we will be in the iterative step that makes different choices. This unified approach between the KAM theorem and exponentially small estimates appears also in [DG1].

4.1. Preliminary estimates and notation. We will be considering area preserving maps f which are defined in a neighborhood of $[-\delta, \delta] \times \mathbb{T}^1$ to itself.

These maps will have the form

$$f(p, q) = (p, q + \omega_0 + \kappa p^M) + O(p^{M+1})$$

for some $\kappa \neq 0$.

By Lemma 2.2, we can find an f_ε in such a way that the $f_0(p, q) = (p, q+\omega_0+\kappa p^M)$. The Hamiltonian of this deformation will be $F_\varepsilon = O(p^{M+1})$.

We will write for these type of families $F_\varepsilon(p, q) = I_\varepsilon(p) + E_\varepsilon(p, q)$, where again I_ε will be thought of as the integrable part. We will denote by i_ε the deformation with initial point f_0 and with Hamiltonian I_ε: $i_\varepsilon(p, q) = \big(p, q + \omega_0 + \kappa p^M + \int_0^\varepsilon ds\, \partial_p I_s(p)\big)$.

We note that these families are a particular case of the families we have considered in section 3. (In particular, $\kappa < 0$ and $M = 2$ for the example (1.1).) In that section, we allowed a dependence in another parameter ω. The families we consider here can be considered as embedded in families depending on ω but such that the dependence on ω is trivial. Clearly, all the results of section 3 that do not rely on the dependence on ω being nontrivial will go through as stated using the elementary device of writing the extra variable ω and noticing that the functions we consider do not depend on ω. We will use this completely elementary device without too much of an explicit mention.

For the purposes of this section, it will be sufficient to use particular cases of the neighborhoods $\Sigma_{\Omega,\alpha,\beta,U}$. Since all the objects we will consider will not depend on ω, we will not need to consider objects that depend on this; in particular we can suppress U from the notation.

We will also introduce the simplified domains

$$\Sigma_\delta = \{(p, q, \varepsilon) \mid |p| \leq \delta, \quad |\Im\, q| \leq \delta, \quad d(\varepsilon, [0, 1)) \leq \delta\}$$

and, given a family of functions $H_\varepsilon(p, q)$, we will denote

$$\|H\|_\delta \equiv \sup_{(p,q,\varepsilon)\in\Sigma_\delta} |H_\varepsilon(p, q)|.$$

Since we will be working with functions that vanish at the origin to a high order, it is worth remarking that Cauchy bounds can be improved for them.

PROPOSITION 4.1. *Let $H_\varepsilon(p, q)$ be such that $H_\varepsilon(p, q) = p^n J_\varepsilon(p, q)$. Then, provided that the norms are defined,*

(i) $\|J\|_\delta = \delta^{-n}\|H\|_\delta$

and, for $\delta' < \delta$, we have

(ii) $\|H\|_{\delta'} \leq (\delta'/\delta)^n\|H\|_\delta$;

(iii) $\|\nabla H\|_{\delta'} \leq (n/\delta' + (\delta - \delta')^{-1})(\delta'/\delta)^n\|H\|_\delta$.

Proof. By the maximum modulus principle

$$\|H\|_\delta = \sup_{\Sigma_\delta} |H_\varepsilon(p, q)| = \sup_{\substack{|p|=\delta \\ |\Im\ q|\leq\delta \\ d(\varepsilon,[0,1])\leq\delta}} |H_\varepsilon(p, q)| = \delta^n \sup_{\Sigma_\delta} |J_\varepsilon(p, q)| = \delta^n\|J\|_\delta.$$

This proves (i). Then,

$$\|H\|_{\delta'} = \delta'^n\|J\|_{\delta'} \leq \delta'^n\|J\|_\delta = (\delta'/\delta)^n\|H\|_\delta.$$

Furthermore,

$$\begin{aligned}\|\nabla(p^n J_\varepsilon)\|_{\delta'} &= \|(np^{n-1}J_\varepsilon + p^n\partial_p J_\varepsilon, p^n\partial_q J_\varepsilon)\|_{\delta'}\\ &\leq n\delta'^{(n-1)}\delta^{-n}\|H\|_\delta + \delta'^n\|\nabla J_\varepsilon\|_{\delta'}\\ &\leq n\delta'^{-1}(\delta'/\delta)^n\|H\|_\delta + \delta'^n(\delta-\delta')^{-1}\|J_\varepsilon\|_\delta\\ &\leq (n\delta'^{-1}+(\delta-\delta')^{-1})(\delta'/\delta)^n\|H\|_\delta. \qquad \square\end{aligned}$$

4.2. Reduction of maps to integrable in a neighborhood of a Diophantine circle. The key step in the proof of Theorem 1.6 is the following. Once we prove this result, the proof will be the same as in [FL].

LEMMA 4.2. *Let ω_0 be a Diophantine number, M an integer. Let f be an analytic area preserving map of the form*

$$f(p,q) = \left(p, q+\omega_0+\kappa p^M\right) + O\left(p^{M+1}\right)$$

for some $\kappa \neq 0$. Then,

(i) *for every $N \in \natural$ we can find an analytic canonical transformation such that*

$$g_N^{-1}\circ T\circ g_N(p,q) = (p, q+\Omega_N(p)) + R_N(p,q) \tag{4.1}$$

with Ω_N analytic, $\Omega_N(p) = \omega_0 + \kappa p^M + O(p^{M+1})$, and $|R_N(p,q)| \leq C_N|p|^N$.

(ii) *Moreover, we can find $\mu_1, \mu_2 > 0$ depending only on M and the Diophantine properties of ω_0, such that for sufficiently small δ, choosing $N = K\delta^{\mu_1}$, we have*

$$\|R_N\|_\delta \leq K\exp(-K^{-1}\delta^{-\mu_2}). \tag{4.2}$$

Remark. We note that Lemma 4.2, besides giving some control on the periodic orbits that we will use to prove Theorem 1.6, also provides control over other orbits. Notably, it shows that critical circles are approximated by KAM circles. Indeed, the density of KAM circles in a neighborhood of size δ of a critical circle will be bigger than $1 - C_1\exp(-C_2\delta^{-\alpha})$ for some positive C_1, C_2, α.

Remark. We observe that the first part of the claim, the reduction to an integrable form could go through with less differentiability. If we only want that $g_N \in \mathcal{C}^4$ (which we will show is enough to show that the residue goes to zero faster than $|\omega_0 - m/n|^{N/M}$) it would suffice to assume that f is $\mathcal{C}^r$ with r depending on N and the Diophantine properties of ω_0. Of course, the quantitative estimates (4.2) depend on the analyticity properties. The first part of the claim is much easier to prove, since, as we will see, it only entails matching powers of p in an equation that expresses the desired result. We note that this is enough to show using the methods that we will develop later that if there is a finitely differentiable circle, then the residue of a periodic orbit of type m/n is smaller than a power of $|\omega_0 - m/n|$. This power can be made as large as we want by assuming that the differentiability is high enough.

Proof of (i). If we denote by $f_0(p,q) = (p, q+\omega_0+\kappa p^M)$, by Lemma 2.2 we can find an analytic family f_ε that interpolates between f_0 and f. The Hamiltonian of this family F_ε^0 will be an analytic function of (p,q,ε) in a complex neighborhood of Σ_δ.

To prove that we can find g_N so that (4.1) holds, we proceed by induction in N and assume that for some $N \geq 2$ we can write our Hamiltonian as

$$F_\varepsilon^N(p,q) = I_\varepsilon^N(p) + E_\varepsilon^N(p,q) \tag{4.3}$$

with

$$E_\varepsilon^N(p,q) = p^N R_\varepsilon^N(p,q).$$

We seek Hamiltonians $G_\varepsilon^N(p,q) = p^N S_\varepsilon^N(q)$ determined in such a way that the family g_ε^N with this Hamiltonian and starting in the identity is such that

$$j_\varepsilon = (g_\varepsilon^N)^{-1} \circ f_\varepsilon^N \circ g_\varepsilon^N \tag{4.4}$$

has a Hamiltonian which is integrable up to a higher order error in p.

We note that

$$g_\varepsilon^N(p,q) = \left(p + p^N \Delta_p(p,q), q + p^{N-1}\Delta_q(p,q)\right), \tag{4.5}$$

where Δ_p, Δ_q are analytic functions. Therefore, the compositions needed to define j_ε in (4.4) make sense in a sufficiently small neighborhood of the circle.

From Proposition 2.1 and (4.5), we can compute the Hamiltonian of j_ε

$$J_\varepsilon = I_\varepsilon^N \circ g_\varepsilon^N + (p + p^N\Delta_p)^N \cdot R_\varepsilon^N \circ g_\varepsilon - G_\varepsilon \circ g_\varepsilon^N + G_\varepsilon^N \circ (f_\varepsilon^N)^{-1} \circ g_\varepsilon^N. \tag{4.6}$$

Expanding the above formula and denoting $R_\varepsilon^N(p,q) = \sum_{i\geq 0} p^i R_\varepsilon^{N,i}(q)$—and analogously for other functions—we obtain

$$\begin{aligned} J_\varepsilon(p,q) \quad = \quad & I_\varepsilon^N(p) + p^N \overline{R_\varepsilon^{N,0}} \\ & + p^N \left\{ \left(R_\varepsilon^{N,0}(q) - \overline{R_\varepsilon^{N,0}} \right) - S_\varepsilon^N(q) + S_\varepsilon^N(q-\omega_0) \right\} + O(p^{N+1}). \end{aligned}$$

Using Lemma 3.2 we now that we can find an analytic S^N so that the term in braces is zero in the domain where the function is defined, which includes a strip around the torus. By the form of the functions, all the compositions needed to define j_ε will be defined in a sufficiently small strip around of the torus.

This establishes the first part of the claim, the fact that we can reduce to any order.

Remark. Rather than using an inductive argument, as we have done, it is possible to show that (4.1) holds to all orders by matching terms in (4.6). We note that the terms of order p^{N+m} have the form

$$R_\varepsilon^{N,m}(q) - S_\varepsilon^{N,m}(q) + S_\varepsilon^{N,m}(q-\omega_0) + \tilde{R}_\varepsilon^{N,m-1}(p,q),$$

where $\tilde{R}_\varepsilon^{N,m-1}$ is a polynomial expression in $R_\varepsilon^{N,i}$, $S_\varepsilon^{N,i}$, $i \leq m-1$ and their derivatives and the derivatives of I. Again, we can use Lemma 3.2 to prove that a solution exists to all orders in p^n.

This method clearly shows that the coefficients of the expansion in the reduction are uniquely determined by the map and the torus, and are independent of the procedure. For example, in [OS], a different procedure using generating functions is used for twist maps and one can find the remark that the coefficients of this normal form are unique. (For the situation we are considering here, generating functions are not so convenient since the mixed variables are not a good system of coordinates in a neighborhood of the invariant torus. Nevertheless, the formalism that we developed above allows us to reach the same uniqueness conclusions.)

To obtain the estimates on the remainders of the reduction, we use a slightly different procedure. We use (3.20) and determine G_ε in exactly the same way as in section 3.

We can apply Lemma 3.6—which does not depend on Ω being nondegenerate—to obtain, with the notation introduced there, (3.53) and (3.54) provided that the inductive hypothesis hold.

LEMMA 4.3. *Let Ω be the frequency function* (3.11) *defined in $\Sigma_{\Omega,\alpha,\beta,U}$ as in* (3.9)*. Let Δ be defined as in* (3.23) *and let σ be a positive number. Assume that* (3.17), (3.28), (3.33)*, and* (3.48) *hold. Consider $\tilde{\Omega}$, the new frequency function defined by*

$$\tilde{\Omega}(\omega,\varepsilon,p) = \Omega(\omega,\varepsilon,p) + \int_0^\varepsilon ds \, \frac{\partial}{\partial p}\Delta(\omega,s,p) \ . \tag{4.7}$$

Then, for any $\tilde{\alpha} \le \alpha$ satisfying

$$K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U} \le \tilde{\alpha} \ , \tag{4.8}$$

we have

(i) $\|\Omega - \tilde{\Omega}\|_{U_\sigma} \le K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U} \le \tilde{\alpha}$;

(ii) *for $\tilde{\alpha}$ as before, $\tilde{\beta} = \beta - \alpha - 4\sigma$, $\tilde{U} \equiv U_{4\sigma}$, we have*

$$\Sigma_{\tilde{\Omega},\tilde{\alpha},\tilde{\beta},\tilde{U}} \subset \Sigma_{\Omega,2\alpha,\beta-4\sigma,U_{4\sigma}};$$

(iii)

$$\begin{aligned} \left\| \left(\tfrac{\partial^M}{\partial p^M}\tilde{\Omega}\right)^{-1} \right\|_{\tilde{U}} &\le \left\| \left(\tfrac{\partial^M}{\partial p^M}\tilde{\Omega}\right)^{-1} \right\|_{U_{4\sigma}} \\ &\le \left\| \left(\tfrac{\partial^M}{\partial p^M}\Omega\right)^{-1} \right\|_{U} + K\sigma^{-M-1}\|E\|_{\Omega,\alpha,\beta,U}; \end{aligned}$$

(iv) *the inequalities* (3.18) *hold, that is, for $\tau = 2\nu + 3$*

$$\|\tilde{E}\|_{\tilde{\Omega},\tilde{\alpha},\tilde{\beta},\tilde{U}} \le K\sigma^{-\tau}\|E\|_{\Omega,\alpha,\beta,U}(\|E\|_{\Omega,\alpha,\beta,U} + \tilde{\alpha}).$$

The only difference between the proofs of Lemma 3.7 and Lemma 4.3 is that in Lemma 4.3 we do not need to worry about the nondegeneracy in Ω with respect to ω. Item (iii) in Lemma 4.3 is just an slight generalization of the standard implicit function theorem. □

We also note that if E is $O(p^L)$, then G is also $O(p^L)$ and, as a consequence, all the terms in the decomposition of $\tilde{E}$ according to (3.20) are $O(p^{2L-1})$ except $(G_{\omega,\varepsilon} \circ i_{\omega,\varepsilon}^{-1} - G_{\omega,\varepsilon} \circ T^0)$ which is only $O(p^{L+M})$. We note that, for high enough L, $2L-1 > L+M$ so that, for large enough L the order of tangency grows by M in each step.

We can therefore assume that if n steps, the resulting nonintegrable part is $O(p^{Mn-A})$ where A is a number that may depend only on M and not on n. The number A takes into account that in the first steps of the iteration it could happen that $2L-1$ is smaller than $L+M$.

Remark. One could have obtained slightly more sophisticated estimates taking advantage of the fact that the functions we are considering vanish with powers of p and we can use the sharper Proposition 4.1 instead of Lemma 3.1. As it turns out, this does not make an appreciable difference in the final answer and it would require that the estimates leading to Lemma 3.6 are redone.

Proof of part (ii) *of Lemma* 4.3: *Iteration of the inductive step.* Now we discuss the possibility and the effect of iterating the inductive step. Since the goals are quite different than in the iteration leading to the KAM theorem, the choices that we will make in domain losses etc. will also be quite different. In our case, we are not interested in having some analyticity domain left (the existence of an analytic torus is part of the assumptions); rather, we are interested in obtaining control of the remainders in a wide domain.

We will take

$$\alpha_{n+1} = cn^{-\eta}, \qquad \beta_n = cn^{-\gamma} \tag{4.9}$$

with η, $\gamma > 0$ chosen in such a way that

$$(\gamma + 1)\tau - \eta < 0. \tag{4.10}$$

Note that then, $\eta > \tau\gamma \geq \gamma + 1$ so that the domains in the p variable are smaller than those in the q variable. Moreover, $\sigma_n = (\beta_n - \beta_{n+1})/4 = c\gamma n^{-\gamma-1} + O(n^{-\gamma-2})$ and we can bound $\sigma_n^{-\tau} \leq Kn^{\tau(\gamma+1)}$. Note also that it also follows from (4.10) that $\eta > \tau$ and that given any $\eta > \tau$ we can chose $\gamma > 0$ in such a way that (4.10) is satisfied.

We claim that if the iterative step can be iterated N times, and c as in (4.9) is sufficiently small, we have

$$\|E^N\|_{\Omega^N,\alpha_N,\beta_N,U^N} \leq (N!)^{(\gamma+1)\tau-\eta}. \tag{4.11}$$

We can proceed by induction. Note that if (4.11) were true, we could, for $N > N_0(c)$, obtain the bound $\|E^N\|_{\Omega^N,\alpha_N,\beta_N,U^N} + \alpha_{N+1} \leq K\alpha_{N+1}$. Then,

$$\|E^{N+1}\|_{\Omega^{N+1},\alpha_{N+1},\beta_{N+1},U^{N+1}} \leq (N!)^{(\gamma+1)\tau-\eta} Kc(N+1)^{(\gamma+1)\tau-\eta}$$

which implies the result for $N+1$ when c is small enough.

We note, as in the proof of the KAM theorem, that all but one of the hypotheses of the iterative step are satisfied provided that $\|E^n\|$ is much smaller than σ_n to a fixed power. The only condition that involves the α is (4.8). Namely,

$$K\sigma^{-1}\|E\|_{\Omega,\alpha,\beta,U} \leq \tilde{\alpha}.$$

We note that, if we fix c, we have the hypotheses satisfied for $N > N_1 > N_0$. If we assume that the error is sufficiently small to start with—which can be assumed if we start in a neighborhood sufficiently small—then, we can perform the N_1 steps and then, the iteration can continue. Therefore, if the initial error $\|E^0\|$ is sufficiently small, we can iterate indefinitely. Notice that since E^0 vanishes up to order M in p it suffices to choose c sufficiently small.

Moreover, the estimates (iii) of Lemma 4.3 tell us that we can bound from below $|(\frac{\partial^M}{\partial p^M}\Omega)^{-1}|$ independently of the number of iterates. Then, the domain Σ_δ is contained in all the domains of the form $\Sigma_{\Omega^N,K\delta^{1/M},K\delta^{1/M},U^N}$ provided that U^N contains a neighborhood of the map.

With the choices of α_n β_n that we have made above in (4.9), we see that we can repeat the iterative step described in Lemma 4.3 and obtain control in a 2δ neighborhood of the circle while $cN^{-\eta} \geq K\delta^{1/M}$. That is, $N \leq K^{-1}\delta^{-1/(M\eta)}$.

As we have seen in (4.11), for N large enough—which is implied by δ small enough—we have

$$\|E^N\|_{\Omega^N,\alpha_N,\beta_N,U^N} \leq (N!)^{(\gamma+1)\tau-\eta} \leq \exp\left(-K^{-1}\delta^{-1/(M\eta)}|\log(\delta^{-1/(M\eta)})| + K\right).$$

By worsening slightly the power of δ in the first term, we can suppress the logarithm to simplify the expression

$$\|E^N\|_{\Omega^N,\alpha_N,\beta_N,U^N} \le (N!)^{(\gamma+1)\tau-\eta} \le \exp\left(-K^{-1}\delta^{-1/(M\eta)-\zeta}\right) \tag{4.12}$$

for some small $\zeta > 0$. Now, we note that the system $f_{\omega,\varepsilon}$ is obtained by solving up to time 1 the system

$$\frac{d}{d\varepsilon}x = \mathcal{F}^N_\varepsilon(x) = \mathcal{I}^N_\varepsilon(x) + \mathcal{E}^N_\varepsilon(x).$$

Applying Cauchy bounds to (4.12), we can obtain bounds for $\mathcal{E}^N$ in a δ neighborhood of the origin which are of the same form as (4.12) with an slightly bigger ζ and some bigger K.

Note that, by definition, $\mathcal{I}^N$ generates an integrable flow. Hence, applying the usual estimates for the dependence of the solutions on the vector field, we obtain the result claimed in Lemma 4.2.

Note that the argument we have given shows that we can take $\mu_1 = 1/M(\eta)$ and μ_2 any number strictly smaller. Since we only needed $(\gamma+1)\tau - \eta < 0$, we can choose η any number bigger than τ and then choose γ. Of course, the constants will be worse. $\square$

4.3. Proof of Theorem 1.6 using Lemma 4.2. A possible proof can be made following the argument in [FL].

We note that Theorem 1.6 makes statements about the trace of derivatives of F^n at fixed points of F^n. Since the trace of the derivative of a map at a fixed point is invariant under changes of coordinates, we can study the derivatives of this map in the coordinates provided by Lemma 4.2.

First, we need to obtain some idea of where the periodic orbits could be. We will need to show that if $|\omega_0 - m/n|$ is small, then the orbit is very close to the invariant circle so that, in the coordinates provided by Lemma 4.2, the orbit is close to being the orbit of an integrable system. Note that for the orbit of an integrable system, the derivative is upper triangular with a diagonal which is the identity (hence, for an integrable system the trace of the derivative is 2 and the residue is 0). A second part of the argument is a perturbation argument that shows that if the system is close to integrable, the trace of the derivative is close to 2 and, hence, the residue is small.

The first part of the argument is accomplished by the following proposition.

PROPOSITION 4.4. *For m/n sufficiently close to ω_0, any orbit of type m/n should be contained in annuli of radii $r \pm O(r^{1+\varepsilon})$ where r satisfies $\omega_0 + \kappa_M r^M = m/n$.*

We see that, when M is odd, we find one such r, namely $r = ((\omega_0 - m/n)/\kappa_M)^{1/M}$. When M is even, if $(\omega_0 - m/n)/\kappa_M$ is positive we can find two such r, namely $r = \pm\,(m/n - \omega_0)/\kappa_M)^{1/M}$ and when $(m/n - \omega_0)/\kappa_M$ is negative, we can find none. (In general, for each of the values of r that guarantee the existence of periodic orbits, they will appear in pairs: elliptic and hyperbolic.)

The argument will also show that, when we cannot find any r solving the equation, there are no periodic orbits of type m/n in a sufficiently small neighborhood of the nondegenerate circle.

Proof. If we apply the first claim of Lemma 4.2 to order $2M+2$, we obtain that, in an appropriate system of coordinates, our map can be written as

$$(p,q) \mapsto (p, q + \Omega(p)) + O\left(p^{2M+2}\right) \tag{4.13}$$

with $\Omega(p) = \kappa p^M + O(p^{M+1})$.

In the set $I = [(9/10)r, (11/10)r] \times \mathbb{T}^1$, the mapping (4.13) can be considered as a perturbation of an integrable system.

We note that the frequencies present in the integrable system in the domain considered are

$$\omega_0 + \kappa r^M[(9/10)^M, (11/10)^M] + O(r^{M+1}).$$

Note also that $\frac{d\Omega}{dp} > \kappa M r^{M-1} + O(r^M)$. This lower bound on the derivative is called the twist constant.

We recall that, by standard arguments in Diophantine approximation, we can find ω^* such that $\forall\ i \in \mathbb{Z} \forall\ j \in \mathbb{Z}$, $|\omega^* - i/j|^{-1} \leq Cj^{5/4}$ in any interval of length bigger than KC^{-1}. (It suffices to fix i, j and consider the length of the interval of ω for which the desired inequality fails. See, e.g., [AA, p. 252].)

Hence, we can find two frequencies $\omega_\pm$ such that

(a) they are Diophantine with exponent $\theta - 1 = 5/4$ and with constant $C = r^{-M}$;
(b) $\omega_- < m/n < \omega_+$;
(c) $\omega_+ - \omega_- \leq Kr^M$.

We now recall the quantitative version of the twist mapping theorem [He] that states that if we perturb an integrable system with twist constant σ defined in a range of A of diameter D, by a perturbation of $\mathcal{C}^4$ size ρ, the invariant circles corresponding to a Diophantine frequency of constant C persist provided that $C^2\rho\ \rho\sigma^{-1}/D$ are sufficiently small. Moreover, $\mathcal{C}^1$ distance of these invariant tori to the unperturbed ones can be bound by $\rho\sigma^{-1}$.

If we apply this to the circles of frequencies $\omega_\pm$ in the domain indicated, we see that $\rho = O(r^{2M+2})$, $C = O(R^{-M})$, $\sigma^{-1} = O(r^{-M+1})$, and $D \geq 2/10r$.

Hence, we conclude that these circles with frequency $\omega_\pm$ persist. Since in a sufficiently small neighborhood of the invariant circle, the map is a twist map, all the orbits with rotation number in $[\omega_-, \omega_+]$ have to be contained in the annulus bounded by these two invariant circles, in particular those of rotation number m/n.

This finishes the proof of Proposition 4.4. $\square$

For the cases where we can find an r such that the rotation number of the integrable part is m/n, we can apply Lemma 4.2 with $\delta = 2r$ with r as above to obtain that $\|R_N\|_\delta$ vanishes to order $K^{-1}|\omega - m/n|^{-\mu_1/M}$ and has size smaller than $K\exp(-K^{-1}|\omega - m/n|^{-\mu_2/M})$.

The improved Cauchy estimates, Proposition 4.1, give us that the entries on the matrix DR are smaller than

$$\begin{aligned} &2^{-K^{-1}|\omega_0 - m/n|^{-\mu_1/M}} K\exp\left(-K^{-1}|\omega_0 - m/n|^{-\mu_2/M}\right) \\ &\qquad \leq K\exp\left(-K^{-1}|\omega_0 - m/n|^{-\mu_3}\right) \end{aligned} \tag{4.14}$$

for some $\mu_3 > 0$.

We also note that the derivatives of the integrable part are of the form $DI = \left(\begin{smallmatrix} 1 & a \\ 0 & 1 \end{smallmatrix}\right)$ with a bounded independently of the number of iterates that we need to take in Lemma 4.2.

If we have a periodic orbit of type m/n, by the chain rule we have $DF^n(x) = DF(x_{n-1}) \cdots DF(x)$, where $x_i = F^i(x)$. Note that $DF(x_i) = DI(x_i) + DR(x_i)$.

Therefore, we can apply the following lemma, which appears as Lemma 3.4 of [FL].

Lemma 4.5. *Let* $\{A_i\}_{i=1}^N$ *be a set of* 2×2 *matrices of the form* $A_i = \left(\begin{smallmatrix} 1 & a_i \\ 0 & 1 \end{smallmatrix}\right)$ *with* $\sup_{1 \leq i \leq N} |a_i| \leq A$.

Let $\{B_i\}_{i=1}^N$ satisfy

$$\sup_{\substack{1\le i\le N\\ j,k=1,2}} |(B_i)_{jk} - (A_i)_{jk}| \le \varepsilon \quad \text{with } \varepsilon \le A\ .$$

Then $B = B_1 \cdots B_N$ satisfies

$$|\operatorname{Tr} B - 2| \le 2\left[\left(1 + 3\sqrt{A}\,\sqrt{\varepsilon}\,\right)^N - 1\right].$$

Applying Lemma 4.5 with $A_i = DI(x_i)$, $B_i = DF(x_i)$, we obtain that for sufficiently large n, recalling that Theorem 1.6 includes in the assumptions that $|\omega_0 - m/n| \le 1/n$ and that therefore $K\exp(-K^{-1}|\omega_0 - m/n|^{\mu_3})$ tends to zero

$$\begin{aligned} |\operatorname{Tr} DF^n(x) - 2| &\le 2\left[\left(1 + K\exp(-K^{-1}|\omega_0 - m/n|^{\mu_3})\right)^n - 1\right] \\ &\le nK\exp(-K^{-1}|\omega_0 - m/n|^{\mu_3}) \le K\exp(-K^{-1}|\omega_0 - m/n|^{\mu_4}). \end{aligned}$$

This concludes the proof of Theorem 1.6. □

We also remark that the argument that we gave to locate the periodic orbits also shows that if we have a nondegenerate critical circle, then it is approximated by periodic orbits.

In the cases that we can find an approximate r (i.e., in the case of odd M or, when M is even, that the sign of $\omega_0 - m/n$ is chosen correctly) we see that we can apply the classical Poincaré last geometric theorem [Fr] to $F^n - (0,m)$ and obtain that there are two fixed points of different index, and hence two different periodic orbits of F.

In the case that M is even and the signs are right, since we can find two rings we can obtain four periodic orbits.

Remark. Note that in order to obtain two periodic orbits using this argument, we need to use the modern version of the Poincaré theorem which includes information about the index of the fixed points of $F^n - (0,m)$. The classical Poincaré fixed point theorem (see, e.g., [St, p. 195]) does not provide information about the index and hence, we cannot exclude that the two fixed points of $F^n - (0,m)$ produced by it are part of the same orbit for F.

Remark. In our case, noting that our maps admit a generating function, we could also produce the two periodic orbits using variational methods (see [KH, Theorem 9.3.7].)

A different line of argument that produces quantitative results under stronger hypotheses is the following.

In an annulus $p \in [r - r^{1+\varepsilon}, r + r^{1+\varepsilon}]$ the map is a small perturbation of an integrable map that is nondegenerate. If this perturbation satisfies some nondegeneracy assumptions, one can find two periodic orbits of type m/n. One of them is hyperbolic and another one is elliptic. The first order calculations of these periodic orbits is sometimes called subharmonic Melnikov theory. Formal expansions, including nondegeneracy assumptions that imply that the expansions predict one pair of elliptic and hyperbolic periodic orbits can be found in [Po, sections 74 and 79]. A justification of these expansions for finitely differentiable functions that shows that, under the formal conditions derived in [Po] one can find indeed the periodic orbits with the character predicted by the expansions can be found in [LW, Chapter 2], or in [Po, section 39].

Remark. Note that the above argument only requires estimates about the trace of the derivative. The fact that the trace of the derivative can be studied requires that

$g_N \in \mathcal{C}^1$. The argument that we used to show that, in the coordinates given by g_N, the periodic orbit of period m/n is at a distance not more that $|\omega_0 - m/n|^{1/M}$ requires the twist mapping theorem with Lipschitz estimates and hence that $g_N \in \mathcal{C}^4$. The rest of the argument applying Lemma 4.5 requires only that the map $g_N \in \mathcal{C}^1$. Hence we see that if $g_N \in \mathcal{C}^4$ we have that $|\mathrm{Res}(O_{m,n})| \le K|\omega_0 - m/n|^{N/M}$. Therefore, as we remarked before, to show that the residue goes to zero faster than a power, one needs only finite differentiability and for $\mathcal{C}^\infty$ mappings one can show that the residue goes to zero faster than any power.

Remark. For a Diophantine number (1.3), it holds that $|\omega_0 - p_n/q_n|^{-1} \ge Cq_n^\tau$ for some $\tau \ge 2$, if we take p_n/q_n to be the convergents of the continued fraction expansion of ω_0. Hence, the conclusion of Theorem 1.6 can be written as

$$\mathrm{Res}(O_n) \le C_1 \exp(-C_2 q_n^{\mu'}).$$

Remark. A followup paper [CGM2] of [CGM1] goes on to find scaling relations for the invariant circles with rotation number $\omega_0 = (\sqrt{5}-1)/2$ of $T_{\omega(\varepsilon),\varepsilon}$ as ε goes to a critical value where they cease to exist. These scaling relations suggest that there is a renormalization group description of these invariant circles with the KAM circles corresponding to a trivial fixed point. If this was the case (to our knowledge nobody has yet worked out a precise formulation and computed the trivial fixed points), the residue of a periodic orbit of type F_n/F_{n+1} would go to zero superexponentially fast in n, since for the Fibonacci numbers $F_0 = F_1 = 1$, $F_{n+1} = F_n + F_{n-1}$, one has $|\omega_0 - F_n/F_{n+1}|^{-1} \approx C\omega_0^{-2n}$.

Remark. In higher dimensions, under the nondegeneracy hypotheses of the KAM theorem—which are weaker than twist hypothesis—an argument similar to the one given above has been developed in [T1]. The reduction to integrable normal form up to a very small error can be carried out. Similarly, there is an analogue of Lemma 4.5 that shows that products of sufficiently small perturbations of Jordan blocks with identity in the diagonal, still have characteristic polynomials close to $(t-1)^{2d}$. Therefore, if there is a periodic orbit in a neighborhood of the torus, not only the trace but all the other coefficients of the characteristic polynomial have to converge to those of the Jordan normal form. One important element from our present argument that does not generalize to higher dimensions is the application of the twist mapping theorem to conclude that the distance of the periodic orbits to the invariant circle is bounded by the difference of the rotation numbers. Nevertheless, it is possible to show that if there is an invariant torus, there are periodic orbits that approximate it well and that the characteristic polynomial of the derivative converges to $(t-1)^{2d}$. It has been argued—and implemented numerically in [T2]—that this convergence of the coefficients of the characteristic polynomial of the derivative can be considered as a test of the presence of a KAM torus.

We think that it should be possible to extend the methods presented here to establish one of the implications of Greene's criterion for some invariant torus that satisfy some hypothesis of nondegeneracy weaker than the twist hypothesis.

Acknowledgments. We thank D. del Castillo-Negrete and P. Morrison for many discussions about their work and for encouragement. We also thank A. Haro and C. Simó for helpful discussions and for making their unpublished results available to us. We also thank TICAM, UPC, and IMA for invitations that made this collaboration possible.

REFERENCES

[AA] V.I. Arnol'd, A. Avez, *Ergodic Problems of Classical Mechanics*, W.A. Benjamin, New York, Amsterdam, 1968.

[BHS] H.W. Broer, G.B. Huitema, and M.B. Sevryuk, *Quasi-periodic Motions in Families of Dynamical Systems (Order Amidst Chaos)*, Lecture Notes in Math. 1645, Springer-Verlag, Berlin, 1996.

[BLW] A. Banyaga, R. de la Llave, and C.E. Wayne, *Cohomology equations near hyperbolic points and geometric versions of Sternberg linearization theorems*, J. Geom. Anal., 4 (1996), pp. 613–649.

[BST] H. Broer, C. Simó, and J.C. Tatjer, *Towards global models near homoclinic tangencies of dissipative diffeomorphisms*, Nonlinearity, 11 (1998), pp. 667–770.

[C] D. del-Castillo-Negrete, *Dynamics and Transport in Rotating Fluids and Transition to Chaos in Area Preserving Non-twist Maps*, Ph.D. Thesis, University of Texas, Austin, TX, 1995.

[CGM1] D. del-Castillo-Negrete, J.M. Greene, and P.J. Morrison, *Area preserving nontwist maps: Periodic orbits and transition to chaos*, Phys. D, 91 (1996), pp. 1–23.

[CGM2] D. del-Castillo-Negrete, J.M. Greene, and P.J. Morrison, *Renormalization and transition to chaos in area preserving nontwist maps*, Phys. D, 100 (1997), pp. 311–329.

[CM] D. del-Castillo-Negrete and P.J. Morrison, *Chaotic transport by Rossby waves in shear flow*, Phys. Fluids A, 5 (1993), pp. 948–965.

[DG1] A. Delshams and P. Gutiérrez, *Effective stability and KAM theory*, J. Differential Equations, 128 (1996), pp. 415–490.

[DG2] A. Delshams and P. Gutiérrez, *Estimates on invariant tori near an elliptic equilibrium point of a Hamiltonian system*, J. Differential Equations, 131 (1996), pp. 277–303.

[DMS] H.R. Dullin, J.D. Meiss, and D. Sterling, *Generic twistless bifurcations*, Phys. D, 134 (1999), pp. 153–184.

[FL] C. Falcolini and R. de la Llave, *A rigorous partial justification of Greene's criterion*, J. Statist. Phys., 67 (1992), pp. 609–643.

[Fr] J. Franks, *Generalizations of the Poincaré-Birkhoff theorem*, Ann. of Math. (2), 128 (1988), pp. 139-151.

[Gr] J.M. Greene, *A method for determining a stochastic transition*, J. Math. Phys., 20 (1979), pp. 1183–1201.

[Ha1] A. Haro, *The Primitive Function of an Exact Symplectomorphism*, Preprint 99–105. University of Barcelona, Barcelona, Spain, 1999; available online from http://www.ma.utexas.edu/mp_arc/

[Ha2] A. Haro, *Converse KAM theory for monotone positive symplectomorphisms*, Nonlinearity, 12 (1999), pp. 1299–1322.

[He] M.R. Herman, *Sur les courbes invariantes par les difféomorphismes de l'anneau*, Astérisque, 103–104 (1983).

[HH1] J.E. Howard and S.M. Hohs, *Stochasticity and reconnection in Hamiltonian systems*, Phys. Rev. A, 29 (1984), pp. 418–421.

[HH2] J.E. Howard and J. Humphreys, *Nonmonotonic twist maps*, Phys. D, 80 (1995), pp. 62–72.

[JV] A. Jorba and J. Villanueva, *On the normal behaviour of partially elliptic lower-dimensional tori of Hamiltonian systems*, Nonlinearity, 10 (1997), pp. 783–822.

[K] W.T. Kyner, *Rigorous and formal stability of orbits about an oblate planet*, Mem. Amer. Math. Soc., 81 (1968), pp. 1–27.

[KH] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambrige University Press, Cambridge, UK, 1995.

[KMOP1] J. Koiller, R. Markarian, S. Oliffson-Kamphorst, and S. Pinto de Carvalho, *Time-dependent billiards*, Nonlinearity, 8 (1995), pp. 983–1003.

[KMOP2] J. Koiller, R. Markarian, S. Oliffson-Kamphorst, and S. Pinto de Carvalho, *Static and time-dependent perturbations of the classical elliptical billiard*, J. Statist. Phys., 83 (1996), pp. 127–143.

[KO] Y. Katznelson and D.S. Ornstein, *A new method for twist theorems*, J. Anal. Math., 60 (1983), pp. 157–208.

[LMM] R. de la Llave, J.M. Marco, and R. Moriyón, *Canonical perturbation theories of Anosov diffeomorphisms and regularity results for the Livsic cohomology equation*, Ann. of Math. (2), 123 (1986), pp. 537–611.

[Ll] R. de la Llave, *Introduction to KAM theory*, in Computational Physics, World Scientific, River Edge, NJ, 1992, pp. 73–105; also available online from http://www.ma.utexas.edu/mp_arc/.

[LW] R. de la Llave and C.E. Wayne, *Whiskered and Low Dimensional Invariant Tori for Near Integrable Hamiltonian Systems*, preprint, 1989.

[Mat] J. Mather, *Stability of C^∞ mappings* II: *Infinitesimal stability implies stability*, Ann. of Math. (2), 89 (1969), pp. 254–291.

[McK] R.S. McKay, *Renormalisation in Area Preserving Maps*, Ph.D. Thesis, Princeton University, Princeton, NJ, 1982.

[McK2] R.S. McKay, *On Greene's residue criterion*, Nonlinearity, 5 (1992), pp. 161–187.

[Mo1] J. Moser, *On the volume elements of a manifold*, Trans. Amer. Math. Soc., 120 (1965), pp. 286–294.

[OS] A. Olvera and C. Simó, *Normal forms close to invariant circles of twist maps*, in European Conference in Iteration Theory (ECIT 87), C. Alsina, J. Llibre, C. Mira, C. Simó, G. Targonski, and R. Thibault, eds., World Scientific, River Edge, NJ, 1989, pp. 438–443.

[Po] H. Poincaré, *Les méthodes nouvelles de la mécanique céleste*, Gauthier Villars, Paris, 1899.

[PW] A. Perry and S. Wiggins, *KAM tori are very sticky: Rigorous lower bounds on the time to move away from an invariant Lagrangian torus with linear flow*, Phys. D, 71 (1994), pp. 102–121.

[Ru] H. Rüssmann, *On optimal estimates for the solutions of linear difference equations on the circle*, Celestial Mech. Dynam. Astronom., 14 (1976), pp. 33–37.

[Ru2] H. Rüssmann, *On a new proof of Moser's twist mapping theorem*, Celestial Mech. Dynam. Astronom., 14 (1976), pp. 19–31.

[Si] C. Simó, *Invariant curves of analytic perturbed nontwist area preserving maps*, Reg. Chaotic Dyn., 3 (1998), pp. 180-195.

[SM] C.L. Siegel and J. Moser, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, 1971.

[St] S. Sternberg, *Celestial Mechanics, Part* II, W. A. Benjamin, New York, 1969.

[TL] H.I. Levine, *Singularities of differentiable mappings, notes of a course by R. Thom*, in Proceedings of Liverpool Singularities—Symposium I (1969/1970), Lecture Notes in Math. 192, C.T.C. Wall, ed., Springer-Verlag, Berlin, 1971, pp. 1–89.

[T1] S. Tompaidis, *Approximation of invariant surfaces by periodic orbits in high-dimensional maps: Some rigorous results*, Experiment. Math., 5 (1996), pp. 197–209.

[T2] S. Tompaidis, *Numerical study of invariant sets of a quasiperiodic perturbation of a symplectic map*, Experiment. Math., 5 (1996), pp. 211–230.

[VG] T.P. Valkering and S.A. van Gils, *Bifurcation of periodic orbits near a frequency maximum in near-integrable driven oscillators with friction*, Z. Angew. Math. Phys., 44 (1993), pp. 103–130.

[W] A. Weinstein, *Lagrangian submanifolds and Hamiltonian systems*, Ann. of Math. (2), 98 (1973), pp. 377–410.

[ZZSUC] G. Zaslavsky, *Stochastic web and diffusion of particles in a magnetic field*, Soviet Phys. JETP, 64 (1987), pp. 294–303. Translated from Zh. Èksper. Teoret. Fiz., 91, (1986), pp. 500–516.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1270–1294

GLOBAL BEHAVIOR OF THE CAUCHY PROBLEM FOR SOME CRITICAL NONLINEAR PARABOLIC EQUATIONS[*]

J. A. AGUILAR CRESPO[†] AND I. PERAL ALONSO[†]

Abstract. The paper deals with the following nonlinear parabolic problem:

$$\begin{cases} u_t - \Delta_p u &=& \lambda \dfrac{u^{p-1}}{|x|^p} \quad x \in \mathbb{R}^N,\, t > 0, \\ u(x,0) &=& u_0(x) \geq 0, \end{cases}$$

where $1 < p < N$ and $\lambda > 0$. The existence or nonexistence (blow-up) of global solution is analyzed. Also the finite time extinction for solutions in the case $1 < p < 2$ in bounded and unbounded domains is studied. This behavior depends on the relationships between λ, N, p and the integrability of u_0.

Key words. nonlinear parabolic equations, p-laplacian, existence, behavior of solutions, critical problems, Hardy inequality

AMS subject classifications. 35K25, 35K55, 35K57, 35K65

PII. S0036141098341137

1. Introduction.

In this work the starting points are the paper by Baras and Goldstein [1] for the heat equation and the paper [12]. In the case of the heat equation the problem considered appears in a natural way by linearizing reaction-diffusion models with convex *supercritical* nonlinearities at some unbounded, *singular* solution (see [7], [13], [16], [19], and [20]).

In this paper we deal with the nonnegative solutions of the following Cauchy problem:

$$(1.1) \qquad \begin{cases} u_t - \Delta_p u &=& u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = \lambda \dfrac{u^{p-1}}{|x|^p} \quad x \in \mathbb{R}^N,\, t > 0, \\ u(x,0) &=& u_0(x) \geq 0, \end{cases}$$

where $1 < p < N$ and $\lambda > 0$. If $p = 2$ this problem corresponds to the problem studied in the pioneering paper [1].

Our attention will be focused on the role that the potential $|x|^{-p}$ plays. Notice that the integrability is critical since it belongs to L^q_{loc} $\forall q < N/p$ but not for $q = N/p$. Then it is a borderline case in the parabolic quasi-linear theory of regularity and uniqueness (see [8], [18], and [17]).

Before describing the main results in this work, we would like to emphasize the dependence of the results on the relation of p with respect to 2 and the relation of the parameter λ with respect to $\lambda_{N,p} = ((N-p)/p)^p$, inverse of the best constant for the Hardy inequality explained in section 2. (See also [14] and [12] for details on this topic.)

The main contributions in the paper are the following:

(I) For existence results:

(i) If $\lambda \leq \lambda_{N,p}$, the global existence holds $\forall p$, $1 < p < N$.

[*] Received by the editors June 24, 1998; accepted for publication (in revised form) October 19, 1999; published electronically May 15, 2000. This work was partially supported by D.G.E.S. (M.E.C., Spain) project PB97-0052.

http://www.siam.org/journals/sima/31-6/34113.html

[†] Departamento de Matemáticas, Universidad Autónoma de Madrid Cantoblanco, 28049 Madrid, Spain (jaa19@etna.tid.es, ireneo.peral@uam.es).

(ii) If $1 < p < 2N/(N+1)$, the global existence holds $\forall \lambda > 0$.

(iii) If $2N/(N+1) < p < N$ and $\lambda > \lambda_{N,p}$ there is no solution. The case $p = 2$ is the previous result in [1] and the case $p > 2$ is obtained in [12].

(II) For the behavior of solutions:

(i) For the Dirichlet problem in a bounded domain and if $2N/(N+1) < p < 2$, $0 < \lambda < \lambda_{N,p}$ there exists $T^* = T^*(u_0, \Omega) < \infty$, finite time of extinction.

(ii) As above for the Dirichlet problem in a bounded domain, if $1 < p < 2N/(N+2)$ and $0 < \lambda < \mu_{N,p}$, where

$$\mu_{N,p} = \lambda_{N,p}(s-1)\left(\frac{p}{p+s-2}\right)^p, \quad s = N\left(\frac{2-p}{p}\right),$$

the finite time of extinction $T^* = T^*(u_0) < \infty$ is independent of Ω. As a consequence under these restrictions on p and λ we are able to show finite time of extinction for the Cauchy problem in the whole $\mathbb{R}^N$.

(iii) We also prove that for λ large (in particular $\lambda > \lambda_{N,p}$) no extinction time exists.

We indicate some possible extensions of our results in the last section. Roughly speaking we can say that for $p \geq 2N/(N+1)$, the constant $\lambda_{N,p}$ plays an important role for the existence of solutions while, for $1 < p < 2N/(N+1)$, only the extinction in finite time depends on the Hardy inequality.

We will look for solutions obtained passing to the limit on solutions to problems in bounded domains and with truncated potential. The uniqueness is, in general, not true.

Concerning the problem about the convergence to the initial data, a complete analysis for data in L^1 when $\lambda = 0$ can be seen in [9], [10], and [21]. This kind of problem is out of the scope of this work: we limit ourselves to see that the data are attained in a weak sense. The optimal integrability condition on u_0 and the optimal way in which in general the initial data are attained for a given solution, jointly with a more detailed analysis of the blow-up, will be the subject of a future research. (See also Remark 3.6 at the end of section 3.)

The paper is organized as follows. The next section is devoted to introduce the truncated problems, the Hardy inequality, and the application of some general results by [5] in this context. In section 3 we study the existence of solutions to (1.1). The methods that we use are classical results from [8], [11], [18], and [17] in general, and in particular for the case $1 < p < 2$ we also use a self-similar solution to (1.1) with zero initial data, the above mentioned results in [5], and arguments from [4], [3], [8], and [12]. In section 4 we obtain the above quoted results about the finite extinction time if $1 < p < 2$ and λ are small enough. See [2], [8], and [15] for the case $\lambda = 0$. Notice that in our critical problem finite time extinction is a delicate property, depending on the *fast diffusion* due to the fact that $1 < p < 2$ and the reaction on the right-hand side through λ. The results about finite time extinction are (almost) optimal (see section 4). Finally, section 5 contains some remarks about the behavior (existence, uniqueness, and finite time extinction) of the solutions corresponding to more general potentials and nonlinearities. We will mainly consider $1 < p < 2$ and nonlinearities of *asymptotic power type*.

To extend the idea of sublinear and superlinear growth in the case $p = 2$, hereafter we will call the nonlinearity f subdiffusive (respectively, superdiffusive) in $u \to a$, where either $a = 0$ or $a = +\infty$, if the following holds:

$$\lim_{u\to a} \frac{f(u)}{u^{q-1}} = c_1 \quad \text{for some } q, \quad 1 < q < p, \quad (\text{respectively, } p < q).$$

2. The truncated problems and compactness arguments. Consider the problems for $1 < p < N$, $\lambda > 0$, $n > 0$:

$$\text{(2.1)} \qquad \begin{cases} u_{nt} - \Delta_p u_n &= \lambda W_n(x) u_n^{p-1}, & x \in \Omega,\ t > 0, \\ u_n(x,0) &= T_n(u_0(x)), & \\ u_n(x,t) &= 0, & x \in \partial\Omega,\ t > 0, \end{cases}$$

where $W_n(x) = T_n(|x|^{-p})$, T_n is the truncature at height n ($T_n(\zeta) = \min(n, \zeta)$ for $\zeta \geq 0$), $u_0(x) \geq 0$, $u_0 \in L^2(\Omega)$ and Ω is a bounded domain in $\mathbb{R}^N$. These problems are called the *truncated problems* since they are obtained from the following initial boundary problem with zero Dirichlet boundary data on Ω (*untruncated problem*):

$$\text{(2.2)} \qquad \begin{cases} u_t - \Delta_p u &= \lambda \dfrac{u^{p-1}}{|x|^p}, & x \in \Omega,\ t > 0, \\ u(x,0) &= u_0(x), & x \in \Omega, \\ u(x,t) &= 0, & x \in \partial\Omega,\ t > 0 \end{cases}$$

by applying the truncature T_n to both the potential $|x|^{-p}$ and the initial data. It has to be noted that

$$W_n(x) \leq |x|^{-p} \in L^r_{loc}(\mathbb{R}^N) \quad \text{uniformly, for } 1 \leq r < N/p.$$

In the particular case $1 < p < 2$ the existence of bounded weak solution u_n to (2.1) in $Q = \Omega \times (0,T)$ is shown in the following elementary way. The boundedness of the weak solution to the truncated problem will be important in section 4 to get some convenient estimates.

LEMMA 2.1. *Let $1 < p < 2$. For every $n > 0$, there exists a weak solution u_n to the truncated problem* (2.1)*. These solutions verify $0 \leq u_1 \leq u_2 \leq \cdots \leq u_n \leq \cdots$; in addition,*

$$u_n \in L^\infty(0,T; L^\infty(\Omega)) \cap L^p(0,T; W_0^{1,p}(\Omega))$$

for $T > 0$ fixed.

Proof. There exists a supersolution to the truncated problem (2.1), namely

$$\phi_n(t) = (\lambda n(2-p)(t+T_0))^{1/(2-p)}, \quad \text{where } T_0 \geq (\lambda n^{p-1}(2-p))^{-1}.$$

It has to be remarked that this supersolution does not depend on the potential; it depends only on the truncature level. Consider the following problems with $k > 0$ and n fixed ($n \geq 1$):

$$\begin{cases} (v_{n,k})_t - \Delta_p v_{n,k} &= \lambda W_n(x) v_{n,k-1}^{p-1}, & x \in \Omega,\ t > 0, \\ v_{n,k}(x,0) &= T_n(u_0(x)), & x \in \Omega, \\ v_{n,k}(x,t) &= 0, & x \in \partial\Omega,\ t > 0, \end{cases}$$

where $v_{n,0} = \phi_n$; there exists a unique weak solution $v_{n,k} \in C([0,T], L^2(\Omega)) \cap L^p(0,T; W_0^{1,p}(\Omega))$ (see [8] and [18]). Since $v_{n,0} = \phi_n \leq v_{n+1,0} = \phi_{n+1}$ and the potential and initial data are both bounded, a recurrence argument using the weak comparison principle implies $v_{n,k} \le v_{n+1,k}$, $k > 0$, because initially

$$\lambda W_n(x)\phi_n^{p-1}(t) \le \lambda W_{n+1}(x)\phi_{n+1}^{p-1}(t) \quad \text{in } Q.$$

As a consequence, taking the respective limits in each iteration, we conclude that $u_n \le u_{n+1}$. On the other hand, these weak solutions u_n of (2.1) are bounded in Q, for $T > 0$ fixed, since the supersolution ϕ_n is bounded for $T > 0$ fixed. □

In the previous result it is not possible to get a uniform L^∞-estimate.

One of the main tools is the following Hardy inequality that will be used in a systematic way in this paper.

LEMMA 2.2 (Hardy inequality). *If $1 < p < N$ and $u \in W^{1,p}(\mathbb{R}^N)$, then*

$$\int_{\mathbb{R}^N} \frac{|u|^p}{|x|^p}\,dx \le \lambda_{N,p}^{-1} \int_{\mathbb{R}^N} |\nabla u|^p\,dx, \qquad \lambda_{N,p} = \left(\frac{N-p}{p}\right)^p,$$

where $\lambda_{N,p}^{-1}$ is optimal.

A proof of this result can be found in [12]; see also [14] for $N = 1$. Notice that the constant is not attained in $W^{1,p}(\mathbb{R}^N)$.

In general, $\forall p,\ 1 < p < N$, we will use a sequence u_n as approximate solutions to obtain a solution u of (2.2) as the limit of u_n on any fixed bounded domain Ω. To this end, we need some compactness results that allow us to pass to the limit in that sequence. One of these compactness results is shown in [5] for a more general context; we adapt the proof of that result to our case as follows:

Fix $T > 0$ and $\Omega \subset \mathbb{R}^N$, a bounded domain, and let $Q = \Omega \times (0,T)$, $g_n = W_n u_n^{p-1}$. For $1 < p < N$ let us define $p^* = Np/(N-p)$, the critical Sobolev exponent. Let us first show the following lemma.

LEMMA 2.3. *If u_n, $n > 0$ is a sequence uniformly bounded in $L^p(0,T;W^{1,p}(\Omega))$, with $1 < p < N$, then*

(a) *$W_n u_n^{p-1} \in L^{p'}(0,T;W^{-1,p'}(\Omega))$, uniformly.*

(b) *$W_n u_n^{p-1} \in L^q(Q)$ uniformly for $1 \le q < (p^*)' = (1-(N-p)/(Np))^{-1}$.*

Proof. Let $g_n = W_n u_n^{p-1}$. (a) Take $\psi \in L^p(0,T;W^{1,p}(\Omega))$, $\psi \ge 0$; using Hölder and Hardy inequalities

$$\begin{aligned}
\int_0^T \int_\Omega g_n \psi\,dx\,dt &\le \int_0^T \int_\Omega \frac{u_n^{p-1}}{|x|^p}\psi\,dx\,dt = \int_0^T \int_\Omega \frac{u_n^{p-1}}{|x|^{p-1}}\frac{\psi}{|x|}\,dx\,dt \\
&\le \int_0^T \left(\int_\Omega \frac{u_n^p}{|x|^p}\,dx\right)^{(p-1)/p} \left(\int_\Omega \frac{\psi^p}{|x|^p}\,dx\right)^{1/p} dt \\
&\le \lambda_{N,p}^{-1} \left(\int_0^T \int_\Omega |\nabla u_n|^p\,dx\,dt\right)^{(p-1)/p} \left(\int_0^T \int_\Omega |\nabla \psi|^p\,dx\,dt\right)^{1/p} \\
&\le C\lambda_{N,p}^{-1} ||\psi||_{L^p(0,T;W^{1,p}(\Omega))}
\end{aligned}$$

since u_n is uniformly bounded in $L^p(0,T;W^{1,p}(\Omega))$, by hypothesis; this means that g_n is uniformly bounded as an operator in the space $L^{p'}(0,T;W^{-1,p'}(\Omega))$.

(b) If u_n is uniformly bounded in $L^p(0,T;W^{1,p}(\Omega))$, then $u_n \in W^{1,p}(\Omega)$ almost everywhere (a.e.) $t \in (0,T)$. By the Rellich–Kondrachov theorem (see, for instance, [6]), $u_n \in L^r(\Omega)$ a.e. $t \in (0,T)$ for $1 \le r < Np/(N-p)$. On the other hand, for

$q \geq 1$, Hölder inequality implies that

$$\int_0^T \int_\Omega g_n^q \, dx \, dt \leq \int_0^T \left(\int_\Omega u_n^{q(p-1)\alpha} \, dx \right)^{1/\alpha} \left(\int_\Omega |x|^{-pq\alpha/(\alpha-1)} \, dx \right)^{(\alpha-1)/\alpha} dt,$$

where $\alpha > 1$ has to be chosen so that $q(p-1)\alpha < Np/(N-p)$ and $q\alpha/(\alpha-1) < N/p$, i.e.,

$$\frac{N}{N-qp} < \alpha < \frac{N}{N-p}\frac{p}{q(p-1)}.$$

Since

$$\frac{N}{N-qp} = \frac{N}{N-p}\frac{N-p}{N-qp},$$

we can always find such an α if the following holds:

$$\frac{N-p}{N-qp} < \frac{p}{q(p-1)}, \text{ i.e., } q < \frac{Np}{Np-(N-p)} = (p^*)'.$$

In this case, we get

$$\begin{aligned} \int_0^T \int_\Omega g_n^q \, dx \, dt &\leq \int_0^T \left(\int_\Omega u_n^{q(p-1)\alpha} \, dx \right)^{1/\alpha} \left(\int_\Omega |x|^{-pq\alpha/(\alpha-1)} \, dx \right)^{(\alpha-1)/\alpha} dt \\ &\leq C(\Omega, N, p, q) \int_0^T \left(\int_\Omega |\nabla u_n|^p \, dx \right)^{q(p-1)/p} \left(\int_\Omega |x|^{-pq\alpha/(\alpha-1)} \, dx \right)^{(\alpha-1)/\alpha} dt. \end{aligned}$$

Since Ω is a bounded domain and u_n is uniformly bounded in $L^p(0,T;W^{1,p}(\Omega))$, then the two integrals above are finite. Thus, for $1 \leq q < (p^*)'$, $g_n = W_n u_n^{p-1} \in L^q(Q)$ uniformly. □

Now consider u_n the sequence of solutions to problem (2.1) and g_n as in the previous lemma. If we assume that u_n is uniformly bounded in $L^p(0,T;W^{1,p}(\Omega))$, by Lemma 2.3 we get $g_n \in L^{p'}(0,T;W^{-1,p'}(\Omega))$ and g_n is bounded in the space of Radon measures on Q, $\mathcal{M}(Q)$, since $g_n \in L^1(Q)$ uniformly. In addition, in these hypotheses, there exists u such that $u_n \rightharpoonup u$ weakly in $L^p(0,T;W^{1,p}(\Omega))$ as $n \to \infty$ (up to a subsequence; see [6, Thm. III.27]). In this way the results by Boccardo and Murat in [5] apply for (2.1), obtaining that $\nabla u_n \to \nabla u$ strongly in $(L^q(\Omega))^N$, for $q < p$, and so we get a solution to (2.2). More precisely, the result by Boccardo and Murat in this context is as follows.

LEMMA 2.4. *If $\{u_n\}$ is a sequence of solutions to the problems* (2.1)*, $n \in \mathbb{N}$, bounded in $L^p(0,T;W^{1,p}(\Omega))$, then there exists a solution $u \in L^\infty_{loc}(0,T;L^2(\Omega)) \cap L^p(0,T;W^{1,p}(\Omega))$ to the problem* (2.2) *in Ω that satisfies the equation in the sense of distributions. The initial data is attained in the sense that $u \in \mathcal{C}([0,T],W^{-1,p'}(\Omega))$.*

Proof. It suffices to adapt the proof of Theorem 4.3 in [5] to problem (2.1). □

Moreover, in the hypotheses above, Lemma 2.3 implies that there exists a function $g \in L^1(Q)$ such that $g_n \rightharpoonup g$ weakly in $L^1(Q)$ (up to a subsequence) as $n \to \infty$. In this way we can apply Theorem 4.1 in [5] for the solutions to (2.1), obtaining for $l > 0$ fixed that

$$\nabla T_l(u_n) \to \nabla T_l(u) \quad \text{strongly in } (L^p_{loc}(Q))^N.$$

This kind of argument will be used in the next section to show the existence of solution to the Cauchy problem for $\lambda < \lambda_{N,p}$ by means of the Hardy inequality.

However, if $\lambda \geq \lambda_{N,p}$ only for $1 < p < 2N/(N+1)$ it is possible to get a global weaker solution and it is necessary to follow a different strategy to show the existence of such a weaker solution to the Cauchy problem (1.1) that will be studied in subsection 3.3 following the arguments in [12] for the case of a bounded domain.

In order to solve the Cauchy problem in the case $\lambda > \lambda_{N,p}$ and $1 < p < 2N/(N+1)$ the compactness arguments are more involved. We will use as supersolution a convenient shift in time, $\bar{S}$, of the self-similar solution S to the Cauchy problem that we obtain in the next section for λ large. Assume that there exists a sequence v_n $(n > 0)$ of solutions to the following problems:

$$\begin{cases} v_{nt} - \Delta_p v_n &= \lambda W_n(x) \tilde{v}_{n-1}^{p-1}, & x \in \Omega_n,\ t > 0, \\ v_n(x,0) &= u_0(x), & x \in \Omega_n, \\ v_n(x,t) &= 0, & x \in \partial\Omega_n,\ t > 0, \end{cases}$$

where $\tilde{v}_{n-1}^{p-1}$ is the extension by zero of v_{n-1} to Ω_n, an increasing nested sequence of bounded domains, u_0 is a bounded function, and assume that v_n is uniformly bounded above by $\overline{S}$ and $u_0 \leq \overline{S}$, that is, $v_n \leq \overline{S}$ $\forall n \geq 0$. In this case we will show in the next section that it is possible to pass to the limit and obtain a solution to the Cauchy problem for $\lambda > 0$ and $1<p<2N/(N+1)$. This critical value $p_1 = 2N/(N+1)$, as we will see in section 3, corresponds with the integrability range of the self-similar solution S. It has to be remarked that the passage to the limit for $1 < p < 2N/(N+2)$ is made using Lemma 2.4. However, the passage to the limit for $2N/(N+2) \leq p < 2N/(N+1)$ is much more delicate, and needs the following lemmas shown in [12], based on the ideas in [3], [4], and [5].

LEMMA 2.5. *If $2N/(N+2) \leq p < 2N/(N+1)$ and v_n is a sequence of positive functions defined on Ω, a bounded domain, verifying that $v_n \leq \overline{S}$ and*

$$\frac{1}{k} \int_0^T \int_{\{v_n < k\}} |\nabla v_n|^p \, dx \, dt \leq M,$$

then the following estimate in the Marcinkiewitz space $\mathcal{M}^{p_2}$ holds

$$|\{(x,t) \in Q \,:\, |\nabla v_n(x,t)| > h\}| \leq C(p,N,T) h^{-p_2},$$

where $Q = \Omega \times [0,T]$, $1 \leq q < N(2-p)/p$ and $p_2 = pq/(q+1)$.

LEMMA 2.6. *In the hypotheses in Lemma 2.5, we get*

(a) $\nabla v_n \to \nabla v$ *a.e. and in measure.*

(b) $|\nabla v_n|^{p-2} \nabla v_n \to |\nabla v|^{p-2} \nabla v$ *in* L^1.

3. Existence results for the Cauchy problem. In this section we show the existence results for the Cauchy problem (1.1). The proofs are based on the compactness results contained in the previous section. We would like to emphasize the different behavior according with the values of λ: if $p \geq 2$ the role of λ is very important, while the role of the Hardy inequality is less important when $p \to 1$. To be precise we classify this section in some subsections.

3.1. A self-similar solution to the Cauchy problem for $p < 2$. Before introducing the existence results for the Cauchy problem (1.1), we are considering the Cauchy problem with $u_0 \equiv 0$ for $p < 2$ fixed. The existence of a positive self-similar solution to this problem for some values of λ and p has to be remarked, obtaining an example of nonuniqueness since the trivial solution is also a solution to this problem (see [12]). More precisely, we look for positive self-similar solutions to the Cauchy problem in all $\mathbb{R}^N$, namely, for solutions like $S(r,t) = t^\alpha f(t^\beta r)$, where $r = |x|$; hence,

$$S_t = \alpha t^{\alpha-1} f + \beta t^{\alpha+\beta-1} r f',\ S_r = t^{\alpha+\beta} f',\ S_{rr} = t^{\alpha+2\beta} f''.$$

Then necessarily
(1) the similarity exponents satisfy $(\alpha - 1) = \alpha(p-1) + \beta p$;
(2) the corresponding ordinary differential equation in the variable $\xi = t^\beta r$ is

$$\alpha f + \beta \xi f' = (p-1)|f'|^{p-2} f'' + \frac{N-1}{\xi} |f'|^{p-2} f' + \frac{\lambda}{\xi^p} |f|^{p-2} f.$$

If we look for solutions of the form $A|\xi|^\gamma$, $A > 0$, we get

$$(1) \quad \gamma = \frac{-p}{2-p}, \qquad (2) \quad |A|^{p-2} = \frac{\alpha + \beta\gamma}{(p-1)|\gamma|^p + (N-p)|\gamma|^{p-2}\gamma + \lambda}.$$

It has been shown in [12] that (2) makes sense if $\lambda > \lambda_{N,p} = ((N-p)/p)^p$, the inverse of the optimal constant in the Hardy inequality. However, if we assume $\lambda > 0$, the values of λ corresponding to the existence of such a self-similar solution are given by

$$\lambda > \mu_{N,p} = \left(\frac{p}{2-p}\right)^{p-1} \left(N - \frac{p}{2-p}\right).$$

In particular, this self-similar solution takes the form

$$S(x,t) = A(\lambda) \left(\frac{t}{|x|^p}\right)^{1/(2-p)},$$

where $x \in \mathbb{R}^N$ and

$$A(\lambda) = \left(\left(\frac{p}{2-p}\right)^{p-1} (p - N(2-p)) + \lambda(2-p)\right)^{1/(2-p)}.$$

Let us introduce the parameter $s = N(2-p)/p$ ($2-p < s < N$, since $1 < p < N$). Then

$$\mu_{N,p} = \left(\frac{p}{2-p}\right)^p (s-1) \leq \lambda_{N,p} = \left(\frac{p+s-2}{2-p}\right)^p.$$

These two critical values are related as follows:

$$\frac{\mu_{N,p}}{\lambda_{N,p}} = (s-1)\left(\frac{p}{p+s-2}\right)^p.$$

It is important to take into account the following facts:

$$\begin{array}{lll}
\mu_{N,p} < 0 \text{ for} & s < 1, \text{ i.e.,} & 2 > p > 2N/(N+1), \\
\mu_{N,p} = 0 \text{ for} & s = 1, \text{ i.e.,} & p = 2N/(N+1), \\
\mu_{N,p} > 0 \text{ for} & 1 < s < N, \text{ i.e.,} & 1 < p < 2N/(N+1), \text{ and} \\
\mu_{N,p} = \lambda_{N,p} \text{ for} & s = 2, \text{ i.e.,} & p = 2N/(N+2).
\end{array}$$

Moreover, $\mu_{N,p}$ is tangent to $\lambda_{N,p}$ at $s = 2$ or $p = 2N/(N+2)$.

The regularity of S depends only on the value of s and is as follows (cf. [12, sect. 6.2]), where $1 \leq q < s$:

$$\begin{aligned} &S \in W^{1,p}_{loc}(\mathbb{R}^N) \cap L^q_{loc}(\mathbb{R}^N), \quad && 2 < s < N, \text{ i.e.,} \quad && 1 < p < \frac{2N}{N+2}, \\ &S \in L^q_{loc}(\mathbb{R}^N), && 1 < s \leq 2, \text{ i.e.,} && \frac{2N}{N+2} \leq p < \frac{2N}{N+1}. \end{aligned}$$

Remark 3.1.

(i) Notice that if $1 < p < 2N/(N+2)$ the critical Sobolev exponent is less than 2; therefore the local integrability properties of S are better than those given by the Sobolev embedding theorem.

(ii) Note that formally there still exists such a self-similar solution in the range $2 - p < s < 1$, namely, $2N/(N+1) < p < 2$. In particular, S does not belong to L^1 locally; S is a solution to the equation away from the origin in the sense of distributions and a.e., with an *infinite mass* concentrate at $(0,0)$ as initial data. If $2N/(N+1) < p < 2$, then $\mu_{N,p} < 0$, namely, we have the nonempty interval $\lambda \in (\mu_{N,p}, 0]$ and for such values of λ the term $\lambda |x|^{-p} u^{p-1}$ is an absorption term.

(iii) Note that a positive shift in time of S, denoted by $\overline{S}(x,t) = S(x, t+t_0)$, $t_0 > 0$, can be used as a supersolution to the general Cauchy problem (1.1), whenever we can take the shift in time large enough to have $u_0 \leq \overline{S}$. □

3.2. Existence results for $\boldsymbol{\lambda < \lambda_{N,p}}$, $\boldsymbol{1 < p < N}$. Now consider the Cauchy problem (1.1) where $1 < p < N$ and $0 < \lambda < \lambda_{N,p}$. We are going to construct a solution to (1.1) as a limit of solutions of approximate problems in bounded domains, where solution in this case means a function

$$u \in L^\infty(0, \infty; L^2_{loc}(\mathbb{R}^N)) \cap L^p(0, T; W^{1,p}_{loc}(\mathbb{R}^N)) \quad \forall T > 0$$

that verifies the equation in (1.1) in the sense of distributions.

In particular, we are assuming that there exists a ball B in $\mathbb{R}^N$ centered at the origin such that $u_0 \in \mathcal{C}_0(\overline{B})$, $u_0 \geq 0$ in B.

THEOREM 3.2. *If $1 < p < N$, $0 < \lambda < \lambda_{N,p}$ and $u_0 \in \mathcal{C}_0(\overline{B})$, $u_0 \geq 0$ in B, then there exists a global solution to* (1.1)*, u, which is obtained as the limit of the solutions u_k of the following problems:*

$$\text{(3.1)} \qquad \left\{ \begin{array}{rcll} u_{kt} - \Delta_p u_k & = & \lambda \dfrac{u_k^{p-1}}{|x|^p}, & x \in B_k,\ t > 0, \\ u_k(x,0) & = & u_0(x), & x \in B_k, \\ u_k(x,t) & = & 0, & x \in \partial B_k,\ t > 0, \end{array} \right.$$

where B_k is the ball of radius k in $\mathbb{R}^N$ centered at the origin. Moreover this solution attains the initial data in the sense that

$$\lim_{t \to 0} \int_{\mathbb{R}^N} u(x,t)\phi(x)\,dx = \int_{\mathbb{R}^N} u_0(x)\phi(x)\,dx \ \forall \phi \in W^{1,p}_0(B_R),\ R > 0.$$

Proof. In these hypotheses, fix $T > 0$ and $R > 0$ large enough such that $B \subset B_R$, and let $Q_R = B_R \times (0,T)$. We know the following facts:

(a) By Theorem 4.1 in [12], there exists a global solution $u_k \geq 0$ for (3.1), verifying

$$u_k \in L^\infty(0, \infty; L^2(B_k)) \cap L^p(0, T; W^{1,p}(B_k)) \quad \forall T > 0,$$

and

$$u_{kt} \in L^2((\epsilon,\infty)\times B_k) \quad \forall \epsilon > 0.$$

(b) If we multiply the equation in (3.1) by u_k and integrate by parts on $B_k \supset B_R$ for k large enough, we get by the Hardy inequality

$$\int_{B_k} u_k^2(x,T)\,dx + \gamma \int_0^T \int_{B_k} |\nabla u_k(x,t)|^p\,dx\,dt \le \int_{B_k} u_0^2(x)\,dx, \qquad \gamma > 0.$$

This implies that the sequence u_k is uniformly bounded in $L^p(0,T;W^{1,p}(B_R))$, and, by Lemma 2.4, we get the existence of $u \in L^\infty_{loc}(0,T;L^2(Q_R)) \cap L^p(0,T;W_0^{1,p}(Q_R))$ such that

$$u_t - \Delta_p u = \lambda \frac{u^{p-1}}{|x|^p}$$

in $\mathcal{D}'(Q_R)$. By a classical argument (see, for instance, [18, p. 156]) we get that $u(x,0) = u_0(x)$ in the sense that

$$\lim_{t\to 0} \int_{\mathbb{R}^N} u(x,t)\phi(x)\,dx = \int_{\mathbb{R}^N} u_0(x)\phi(x)\,dx \; \forall \phi \in W_0^{1,p}(B_R).$$

In addition, since $u_k \to u$ strongly in $L^p_{loc}(Q_R)$, $\nabla u_k \to \nabla u$ strongly in $L^q(Q_R)$, $1 \le q < p$, and $\nabla T_l u_k \to \nabla T_l u$ strongly in $L^p_{loc}(Q_R)$ (see section 2), then we obtain $u_k \to u$ strongly in $W^{1,q}_{loc}(Q_R)$, $1 \le q < p$ and $T_l u_k \to T_l u$ strongly in $W^{1,p}_{loc}(Q_R)$. □

Remark 3.3. If we take $u_0 \in L^r(B)$, with $r \ge 2$, the result is also true, since B is a bounded domain. This allows us to take $u_0 \in L^s(B)$, where $s \ge 2$ (that is, $1 < p \le 2N/(N+2)$). Moreover, this result is also true for $u_0 \in L^2(\mathbb{R}^N)$; the proof is obtained in a similar way by taking truncatures of u_0 on every B_k. □

3.3. Existence results for $\lambda > 0$, $1 < p < 2N/(N+1)$. The existence of solution for the Cauchy problem (1.1) with $\lambda > 0$ and $1 < p < 2N/(N+1)$ is shown in this subsection. The proof relies on an iteration process similar to those in [12], using the self-similar solution S corresponding to λ suitably shifted in time as a supersolution.

Let B denote a ball in $\mathbb{R}^N$; we have the following result.

THEOREM 3.4. *If $1 < p < 2N/(N+1)$, $\lambda > 0$ and $u_0 \in \mathcal{C}_0(\overline{B})$, $u_0 \ge 0$ in B, then there exists a solution u to* (1.1) *in the sense of distributions, which is obtained as the limit of the sequence given by the iterations ($n > 0$)*

$$\left\{\begin{array}{rcll} v_{nt} - \Delta_p v_n &=& \lambda W_n(x)\tilde{v}_{n-1}^{p-1}, & x \in B_{n+1},\ t > 0,\\ v_n(x,0) &=& u_0(x), & x \in B_{n+1},\\ v_n(x,t) &=& 0, & x \in \partial B_{n+1},\ t > 0, \end{array}\right.$$

with

$$\left\{\begin{array}{rcll} v_{0t} - \Delta_p v_0 &=& 0, & x \in B_1,\ t > 0,\\ v_0(x,0) &=& u_0(x), & x \in B_1,\\ v_0(x,t) &=& 0, & x \in \partial B_1,\ t > 0, \end{array}\right.$$

where $\tilde{v}_{n-1} = v_{n-1}$ in B_n, $\tilde{v}_{n-1} = 0$ in $\mathbb{R}^N \setminus B_n$, and B_n is the ball of radius n centered at the origin in $\mathbb{R}^N$.

In fact, if $1 < p < 2N/(N+2)$ *then* $u \in L^\infty(0,\infty; L^2_{loc}(\mathbb{R}^N)) \cap L^p(0,T; W^{1,p}_{loc}(\mathbb{R}^N))$. *This solution satisfies the initial data in the sense of distributions.*

Proof. Since $\lambda > 0$, there exists a self-similar supersolution to the Cauchy problem with zero initial data, S. Let $\overline{S}$ denote a shift in time of S so that $u_0 \le \overline{S}$ (u_0 is bounded). Then we have in B_1

$$\begin{array}{rcll} v_{0t} - \Delta_p v_0 = 0 & \le & \lambda \dfrac{\overline{S}^{p-1}}{|x|^p} = \overline{S}_t - \Delta_p \overline{S}, & x \in B_1,\ t > 0, \\ v_0(x,0) = u_0(x) & \le & \overline{S}(x,0), & x \in B_1, \\ v_0(x,t) = 0 & \le & \overline{S}(x,t), & x \in \partial B_1,\ t > 0. \end{array}$$

Then we conclude that $\tilde{v}_1 \le \overline{S}$. Therefore, by recurrence, it is easy to show that $\tilde{v}_n \le \overline{S}\ \forall n > 0$.

Fix $T > 0$ and $R > 0$ large enough such that $B \subset B_{R+1}$ and take a cutoff function $\varphi = \varphi(x) \in \mathcal{C}_0^\infty(B_{R+1})$, $\varphi \equiv 1$ in B_R, $0 \le \varphi \le 1$ and $|\nabla \varphi| \le C$ in $A_R = B_{R+1} \setminus B_R$ (notice that φ does not depend on t). Let $Q_R = B_R \times (0,T)$ and take $n > R+1$ so that $B_{R+1} \subset B_n$.

Case I. Consider $1 < p < 2N/(N+2)$; since $v_n \in W_0^{1,p}(B_n)$ and $\varphi \in \mathcal{C}_0^\infty(B_{R+1})$, then $v_n \varphi^p \in W_0^{1,p}(B_{R+1})$; if we multiply by $v_n \varphi^p$ the equation satisfied by v_n and integrate, we obtain

$$\int_{B_{R+1}} v_{nt} v_n \varphi^p + \int_{B_{R+1}} \langle |\nabla v_n|^{p-2} \nabla v_n, \nabla(v_n \varphi^p) \rangle = \lambda \int_{B_{R+1}} W_n \tilde{v}_{n-1}^{p-1} v_n \varphi^p.$$

Since $\tilde{v}_{n-1} \le \overline{S}$ and $v_n \le \overline{S}$ in B_n, and integrating on the interval $[0,T]$, we get

$$\begin{aligned} &\frac{1}{2} \int_{B_{R+1}} v_n^2(x,T) \varphi^p + \int_0^T \int_{B_{R+1}} |\nabla v_n|^p \varphi^p \\ &\quad + p \int_0^T \int_{A_R} v_n \varphi^{p-1} \langle |\nabla v_n|^{p-2} \nabla v_n, \nabla \varphi \rangle \\ &\quad \le \frac{1}{2} \int_{B_{R+1}} u_0^2 \varphi^p + \lambda \int_0^T \int_{B_{R+1}} W_n \overline{S}^p. \end{aligned}$$

Using Young and Hölder inequalities,

$$\begin{aligned} \frac{1}{2} \quad & \int_{B_{R+1}} v_n^2(x,T) \varphi^p + \int_0^T \int_{B_R} |\nabla v_n|^p \varphi^p + \int_0^T \int_{A_R} |\nabla v_n|^p \varphi^p \\ & \le \frac{1}{2} \int_{B_{R+1}} u_0^2 \varphi^p - p \int_0^T \int_{A_R} v_n \varphi^{p-1} \langle |\nabla v_n|^{p-2} \nabla v_n, \nabla \varphi \rangle + \lambda \int_0^T \int_{B_{R+1}} W_n \overline{S}^p \\ & \le \frac{1}{2} \int_{B_{R+1}} u_0^2 + p \int_0^T \int_{A_R} (|\nabla v_n| \varphi)^{p-1} v_n |\nabla \varphi| + \lambda \int_0^T \int_{B_{R+1}} W_n \overline{S}^p \\ & \le \frac{1}{2} \int_{B_{R+1}} u_0^2 + \int_0^T \int_{A_R} (|\nabla v_n| \varphi)^p + C_1(p) \int_0^T \int_{A_R} v_n^p \\ & + \lambda \int_0^T \left(\int_{B_{R+1}} W_n^{2/(2-p)} \right)^{(2-p)/2} \left(\int_{B_{R+1}} \overline{S}^2 \right)^{p/2}. \end{aligned}$$

Since we can simplify the terms involving $\varphi^p |\nabla v_n|^p$ on A_R, we obtain, again using

Young inequality and the facts $v_n \le \overline{S}$ in B_n and $W_n \le |x|^{-p}$,

$$\begin{aligned}&\frac{1}{2}\int_{B_R} v_n^2(x,T) + \int_0^T\int_{B_R} |\nabla v_n|^p \\ &\quad \le \frac{1}{2}\int_{B_{R+1}} u_0^2 + C_1(p)\int_0^T\int_{A_R} \overline{S}^p \\ &\quad + \lambda\left(\frac{2-p}{2}\int_0^T\int_{B_{R+1}} |x|^{-2p/(2-p)} + \frac{p}{2}\int_0^T\int_{B_{R+1}} \overline{S}^2\right).\end{aligned}$$

If we define

$$\beta(T) = \int_{B_{R+1}} u_0^2 + C_1(p)\int_0^T\int_{A_R} \overline{S}^p + \lambda(2-p)\int_0^T\int_{B_{R+1}} |x|^{-2p/(2-p)} + \lambda p\int_0^T\int_{B_{R+1}} \overline{S}^2,$$

then $\beta(T)$ is uniformly bounded, since $1 < p < 2N/(N+2)$ and $\overline{S} \in L^q_{loc}(\mathbb{R}^N)$ with $q > 2$. Therefore, we get

$$\int_{B_R} v_n^2(x,T) \le \beta(T) \le c(\lambda, p, T, R, N, ||u_0||_2).$$

This gives a uniform bound for the norm of v_n in $L^2(B_R)$; in addition, since

$$\int_0^T\int_{B_R} |\nabla v_n|^p \le \beta(T) \le c(\lambda, p, T, R, N, ||u_0||_2),$$

we have that the sequence v_n is uniformly bounded in the space $L^p(0,T;W^{1,p}(B_R))$. In this way we can pass to the limit by Lemma 2.4, obtaining a positive global solution u to (1.1) in $\mathcal{D}'(Q_R)$ with $\lambda > 0$ and $1 < p < 2N/(N+2)$. Here the initial data is attained in the same sense as in Theorem 3.2 because we have the same kind of uniform estimates.

Case II. Now let $2N/(N+2) \le p < 2N/(N+1)$; in this range, the regularity of $\overline{S}$ (see section 3.1) implies that the sequence v_n converges in L^q to some $v \le \overline{S}$, for $1 \le q < s \le 2$; so we cannot follow the previous argument in order to show the existence result. We introduce the sets

$$C_{R,k} = B_R \cap \{v_n\varphi^p \ge k\}, \quad \overline{C}_{R,k} = B_R \cap \{v_n\varphi^p < k\}, \quad A_{R,k} = \overline{C}_{R+1,k} \setminus \overline{C}_{R,k}.$$

If we multiply by $T_k(v_n\varphi^p)$ the equation satisfied by v_n and integrate on B_{R+1}, we obtain

$$\int_{B_{R+1}} v_{nt}T_k(v_n\varphi^p) + \int_{B_{R+1}} \langle|\nabla v_n|^{p-2}\nabla v_n, \nabla(T_k(v_n\varphi^p))\rangle = \lambda\int_{B_{R+1}} W_n\tilde{v}_{n-1}^{p-1}T_k(v_n\varphi^p),$$

that is (remember that φ does not depend on t),

$$\begin{aligned}&\frac{1}{2}\int_{\overline{C}_{R+1,k}} (v_n^2)_t\varphi^p + k\int_{C_{R+1,k}} v_{nt} + \int_{\overline{C}_{R+1,k}} \langle|\nabla v_n|^{p-2}\nabla v_n, \nabla(v_n\varphi^p)\rangle \\ &= \lambda\int_{B_{R+1}} W_n\tilde{v}_{n-1}^{p-1}T_k(v_n\varphi^p).\end{aligned}$$

Integrating on $[0,T]$, we get (note that $\tilde{v}_{n-1} \le \overline{S}$ and $0 \le \varphi \le 1$)

$$\begin{aligned}&\frac{1}{2}\int_{\overline{C}_{R+1,k}} v_n^2(x,T)\varphi^p + \\ &\int_0^T\int_{\overline{C}_{R+1,k}} \varphi^p|\nabla v_n|^p + p\int_0^T\int_{A_{R,k}} \varphi^{p-1} v_n\langle|\nabla v_n|^{p-2}\nabla v_n, \nabla\varphi\rangle \\ &\le \frac{1}{2}\int_{\overline{C}_{R+1,k}} u_0^2\varphi^p + k\int_{C_{R+1,k}} u_0 + \lambda k\int_0^T\int_{B_{R+1}} W_n\overline{S}^{p-1}.\end{aligned}$$

This last integral is bounded whenever $2N/(N+2) \le p < 2N/(N+1)$ and u_0 is bounded. On the other hand, note that if $v_n < k$, then $v_n\varphi^p < k$. In other words

$$B_R \cap \{v_n\varphi^p < k\} \supset B_R \cap \{v_n < k\}.$$

Thus, using the Young inequality in a similar way to the case $1 < p < 2N/(N+2)$,

$$\begin{aligned}&\frac{1}{2}\int_{B_R\cap\{v_n<k\}} v_n^2(x,T) + \int_0^T\int_{B_R\cap\{v_n<k\}} |\nabla v_n|^p + \int_0^T\int_{A_{R,k}} \varphi^p|\nabla v_n|^p \\ &\quad\le \frac{1}{2}\int_{\overline{C}_{R+1,k}} u_0^2\varphi^p + \lambda k C(R) + p\int_0^T\int_{A_{R,k}} (\varphi|\nabla v_n|)^{p-1}(v_n|\nabla\varphi|) \\ &\quad\le \frac{1}{2}\int_{B_{R+1}} u_0^2 + \lambda k C(R) + \int_0^T\int_{A_{R,k}} (|\nabla v_n|\varphi)^p + C_1(p)\int_0^T\int_{A_{R,k}} v_n^p.\end{aligned}$$

So we can simplify the terms involving $(|\nabla v_n|\varphi)^p$ on $A_{R,k}$ and, using the fact that $v_n \le \overline{S}$ in B_{n+1}, we obtain

$$\begin{aligned}&\frac{1}{2}\int_{B_R\cap\{v_n<k\}} v_n^2(x,T) + \int_0^T\int_{B_R\cap\{v_n<k\}} |\nabla v_n|^p \\ &\quad\le \frac{1}{2}\int_{B_{R+1}} u_0^2 + \lambda k C(R) + C_1(p)\int_0^T\int_{A_{R,k}} \overline{S}^p.\end{aligned}$$

Though $\overline{S} \in L^q_{loc}$ for $1 \le q < s$ in the range $2N/(N+2) \le p < 2N/(N+1)$, it has to be noted that the last integral is finite, since we are integrating on $A_{R,k}$, which does not contain the origin. Then

$$\frac{1}{k}\int_0^T\int_{B_R\cap\{v_n<k\}} |\nabla v_n|^p\,dx\,dt \le M,$$

where M does not depend on n; this inequality allows us to use Lemmas 2.5 and 2.6, obtaining the existence of a solution u to (1.1) in Q_R with $\lambda > 0$ and $2N/(N+2) \le p < 2N/(N+1)$ in the sense of distributions. Moreover, $u \in L^\infty_{loc}((0,\infty), L^q(\mathbb{R}^N))$ and $|\nabla u| \in \mathcal{M}^{p_2}$, where $\mathcal{M}^{p_2}$ is the Marcinkiewitz space and $p_2 = pq/(q+1)$ with $1 \le q < s = N(2-p)/p$.

The argument to see how the initial data is attained is slightly different in this case. We take into account the integrability of the upper bound S. Following the calculations in [21, p. 330] and since $p < 2$ we are able to prove that $|\nabla u|^{p-1} \in L^r_{loc}$ with $1 < r < p/(2p-2)$. Then we can say that

$$\lim_{t\to 0}\int_{\mathbb{R}^N} u(x,t)\phi(x)\,dx = \int_{\mathbb{R}^N} u_0(x)\phi(x)\,dx \ \forall\phi \in W_0^{1,r/(r-1)}(B_R),\ R > 0. \qquad \square$$

3.4. Blow-up for $\lambda > \lambda_{N,p}$, $p > 2$. Following the ideas in [12], we can show the existence of blow-up for any solution to (1.1) for $p > 2$ and $\lambda > \lambda_{N,p}$, with positive initial data. More precisely we have the following result.

THEOREM 3.5. *Consider the problem*

$$\left\{\begin{array}{rcll} u_t - \Delta_p u &=& \dfrac{\lambda}{|x|^p}|u|^{p-2}u, & x \in \mathbb{R}^N, \quad N > p > 2,\, t > 0,\, \lambda > \lambda_{N,p}, \\ u(x,0) &=& f(x), & x \in \mathbb{R}^N, \end{array}\right. \tag{3.2}$$

where $p > 2$, $\lambda > \lambda_{N,p}$, $f \in L^\infty$, $f \geq 0$, and $f > \delta > 0$ in a neighborhood of the origin. Then (3.2) *has no local solution, in the sense that for any $\epsilon > 0$, there exists $r(\epsilon) > 0$ such that $\lim_{n\to\infty} u_n(x,t) = \infty$ if $|x| \leq r(\epsilon)$ and $t \geq \epsilon$, where u_n are the solutions to the problems with the truncated potentials $W_n(x)$.*

See [12] for the proof of a bounded domain and notice that the case $\Omega = \mathbb{R}^N$ is an elementary consequence.

Remark 3.6. Some final comments to this section are in order.

(i) With some modifications of the methods in [9], [10], and [21] and taking into account the precise construction, for the solutions founded above it should be possible to get a best result about the convergence to the initial data (for instance in L^1_{loc}). Notice that the integrability properties of the upper bound S in the interval $2N/(N+2) < p < 2N/(N+1)$ suggest the conjecture that this behavior must be true for a wider class of initial data, namely, this problem is connected with the question of the optimal regularity of the initial data, and one of the difficulties is the nonuniqueness.

(ii) We get *instantaneous and regionally complete blow-up*, in particular in all L^r-norms, if $p > 2$. If $1 < p < 2N/(N+1)$, according to the previous section, we have *instantaneous blow-up* in L^∞, but some norms are finite and this fact allows us to construct weaker global solutions. In the linear case ($p = 2$) the representation of the solutions by the Green's function implies *infinite speed of propagation* and this is the point that makes easy to prove that the blow-up is complete. In our case, if $p > 2$, the speed of propagation is finite if $\lambda = 0$, and the argument to prove that $\lim_{n\to\infty} u_n(x,t) = \infty$ $\forall (x,t) \in \Omega \times (0,\infty)$ must be different. This seems to be an open question. □

4. Finite time extinction.

4.1. Extinction results. Let Ω be a bounded domain in $\mathbb{R}^N$. If $||\cdot||_q$ denotes the norm in the space $L^q(\Omega)$, we have the following results.

PROPOSITION 4.1. *Let Ω be a bounded domain in $\mathbb{R}^N$; if u is a solution of*

$$\left\{\begin{array}{rcll} u_t - \Delta_p u &=& \lambda\dfrac{u^{p-1}}{|x|^p}, & x \in \Omega,\, t > 0, \\ u(x,0) &=& u_0(x) \in L^2(\Omega), \quad u_0 \geq 0, & \\ u(x,t) &=& 0, & x \in \partial\Omega,\, t > 0, \end{array}\right.$$

where $2N/(N+2) \leq p < 2$ and $0 < \lambda < \lambda_{N,p}$, then there exists a finite time $T^\star$ depending on $N, p, \lambda, |\Omega|$ and $||u_0||_2$ such that

$$u(\cdot,t) \equiv 0 \qquad \forall t \geq T^\star.$$

Moreover, $0 < T^\star \leq \gamma_1 ||u_0||_2^{2-p} |\Omega|^{\frac{p}{2}+\frac{p}{N}-1}$ where γ_1 is a positive constant depending only on N, p, λ.

Proof. Notice that $2N/(N+2) \le p < 2$ implies $s = N(2-p)/p \le 2$ and $p^* > 2$. The part (a) of the proof of Theorem 3.2 shows the existence of a solution u to this problem. If we multiply the equation by u and integrate on Ω, we get by the Hardy inequality

$$\frac{1}{2}\frac{d}{dt}||u(x,t)||_2^2 + \int_\Omega |\nabla u(x,t)|^p\,dx = \lambda \int_\Omega \frac{u^p}{|x|^p}\,dx \le \lambda\lambda_{N,p}^{-1}\int_\Omega |\nabla u(x,t)|^p\,dx.$$

Then, if $\lambda < \lambda_{N,p}$, there exists $\gamma = \gamma(N,p,\lambda) > 0$ such that

$$\frac{1}{2}\frac{d}{dt}||u(x,t)||_2^2 + \gamma\int_\Omega |\nabla u(x,t)|^p dx \le 0.$$

Now, if $2N/(N+2) \le p$, then $p^* = Np/(N-p) \ge 2$, and

$$||u(x,t)||_{p^*} \le \gamma_{N,p}||\nabla u(x,t)||_p.$$

Therefore

$$\frac{1}{2}\frac{d}{dt}||u(x,t)||_2^2 + \gamma||u(x,t)||_{p^*}^p \le 0.$$

Hölder inequality implies that

$$\int_\Omega |u(x,t)|^2\,dx \le |\Omega|^{(p^*-2)/p^*}\left(\int_\Omega |u(x,t)|^{p^*}\,dx\right)^{2/p^*}.$$

Thus

$$\frac{1}{2}\frac{d}{dt}||u(x,t)||_2^2 + \gamma|\Omega|^{1-p/2-p/N}||u(x,t)||_2^p \le 0,$$

and we obtain

$$||u(x,T)||_2 \le ||u_0||_2\left(1 - \frac{(2-p)\gamma|\Omega|^{1-p/2-p/N}T}{||u_0||_2^{2-p}}\right)_+^{1/(2-p)}. \qquad \square$$

Proposition 4.2. *Consider $\Omega \subset \mathbb{R}^N$ a bounded domain and let u be the solution of*

$$(4.1) \qquad \begin{cases} u_t - \Delta_p u &= \lambda\dfrac{u^{p-1}}{|x|^p}, & x \in \Omega,\ t > 0,\\ u(x,0) &= u_0(x) \in L^2(\Omega)\cap L^s(\Omega), \quad u_0 \ge 0, & \\ u(x,t) &= 0, & x \in \partial\Omega,\ t > 0,\end{cases}$$

obtained as the limit of the sequence u_n of solutions to the corresponding truncated problem (2.1)*, where $1 < p < 2N/(N+2)$ ($s = N(2-p)/p > 2$) and $0 < \lambda < \mu_{N,p}$. Then there exists a finite time $T^\star$ depending only upon N, p, λ and $||u_0||_s$, such that*

$$u(\cdot,t) \equiv 0 \qquad \forall t \ge T^\star.$$

Moreover, $0 < T^\star \le \gamma_2||u_0||_s^{2-p}$, $1 < p < 2N/(N+2)$ where γ_2 is a positive constant depending only upon N, p, λ.

Proof. Observe that $u_n \in L^\infty(\Omega) \cap W_0^{1,p}(\Omega)$ by Lemma 2.1; then u_n^{s-1} is an admissible test function. By multiplying u_n^{s-1} by the equation satisfied by u_n, we get

$$\begin{aligned}\frac{1}{s}\frac{d}{dt}||u_n(x,t)||_s^s + \mu_{N,p}\lambda_{N,p}^{-1}\int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p\,dx \\ = \lambda\int_\Omega W_n(x)u_n^{p+s-2}(x,t)\,dx.\end{aligned} \tag{4.2}$$

Since

$$\int_\Omega |\nabla u_n^{(p+s-2)/p}|^p = (s-1)\frac{\lambda_{N,p}}{\mu_{N,p}}\int_\Omega u_n^{s-2}|\nabla u_n|^p,$$

we can conclude that $u_n^{(p+s-2)/p} \in W_0^{1,p}(\Omega)$; therefore, using the Hardy inequality in the right-hand side of (4.2), we obtain

$$\begin{aligned}\frac{1}{s}\frac{d}{dt}||u_n(x,t)||_s^s + \mu_{N,p}\lambda_{N,p}^{-1}\int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p dx \\ \le \lambda\lambda_{N,p}^{-1}\int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p\,dx.\end{aligned}$$

If $\lambda < \mu_{N,p}$, then there exists $\gamma = \gamma(N,p,\lambda) > 0$ such that

$$\frac{1}{s}\frac{d}{dt}||u_n(x,t)||_s^s + \gamma\int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p\,dx \le 0.$$

Since, by Sobolev,

$$\int_\Omega u_n^s(x,t)\,dx = \int_\Omega u_n^{p^*(p+s-2)/p}(x,t)\,dx \le \gamma_{N,p}\left(\int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p\,dx\right)^{N/(N-p)},$$

we obtain

$$\frac{1}{s}\frac{d}{dt}||u_n(x,t)||_s^s + \gamma||u_n(x,t)||_s^{p+s-2} \le 0.$$

Therefore we conclude that

$$||u_n(x,T)||_s \le ||u_0||_s\left(1 - \frac{(2-p)\gamma T}{||u_0||_s^{2-p}}\right)_+^{1/(2-p)}.$$

That is, there exists a uniform finite extinction time for any solution of the truncated problem corresponding to (4.1). Since u is obtained as the limit of the nondecreasing (see Lemma 2.1) sequence u_n, then there exists a finite extinction time $T^\star$ for u too, namely,

$$0 < T^\star \le \gamma_2||u_0||_s^{2-p}. \qquad \square$$

Remark 4.3. In the hypotheses in the previous result, we also get that the solution u obtained as the limit of the sequence of solutions to the truncated problems belongs to $L^\infty(0,T^\star;L^s(\Omega))$, $s > 2$. $\square$

Remark 4.4. If $\lambda = 0$, that is, in the homogeneous case, similar results have been shown in [2] (see also [15] and [8, Chap. VIII, sections 2 and 3]) which depend only on N and p (if $\lambda < 0$ one obtains similar results). However, if $\lambda > 0$, then λ must be small enough to obtain finite time extinction. $\square$

Remark 4.5. The following question seems to be an open problem: the existence of finite time extinction for some solutions of (4.1) for $\lambda \in [\mu_{N,p}, \lambda_{N,p}]$ (in bounded domains and in $\mathbb{R}^N$). In some sense, this is equivalent to a uniqueness result in the regularity class L^s, since it is possible to find solutions in $\mathbb{R}^N$ which do not have finite time extinction in this range of λ and do not belong to L^s (see subsection 4.2). On the other hand, a similar estimate can be shown for the finite time extinction of the solutions of the truncated problems, u_n, if $\lambda < \lambda_1(n)\mu_{N,p}/\lambda_{N,p}$, where $\lambda_1(n)$ is the first eigenvalue associated with the following elliptic problem:

$$\left\{\begin{array}{rcll} -\Delta_p \phi_1 & = & \lambda_1(n) W_n(x) \phi_1^{p-1}, & x \in \Omega, \\ \phi_1 & = & 0, & x \in \partial\Omega \end{array}\right.$$

(see subsection 5.3 for more details). Since $\lambda_1(n) \searrow \lambda_{N,p}$ as $n \to \infty$, we obtain again the bound $\lambda < \mu_{N,p}$ for the finite time extinction of the solutions of the untruncated problem. $\square$

We would like to emphasize that the finite extinction time estimate obtained in Proposition 4.2 does not depend on $|\Omega|$ for $\lambda < \mu_{N,p} \le \lambda_{N,p}$ and $s > 2$. This is the main result that we will use in the proof of the finite time extinction for the Cauchy problem with $0 < \lambda < \mu_{N,p}$ and $1 < p < 2N/(N+2)$.

THEOREM 4.6. *If $1 < p < 2N/(N+2)$ $(s > 2)$, $0 < \lambda < \mu_{N,p}$ and $u_0 \in L^2(\mathbb{R}^N) \cap L^s(\mathbb{R}^N)$, there exists a finite time extinction $T^\star$ (depending only upon N, p, λ and u_0) for u, the solution of* (1.1) *obtained in Theorem* 3.2, *such that*

$$u(\cdot, t) \equiv 0 \ \forall t \ge T^\star \text{ and } 0 < T^\star \le \gamma_2 ||u_0||_s^{2-p},$$

where γ_2 is a positive constant depending only upon N, p, λ.

Proof. In these hypotheses, we can use Theorem 3.2 to obtain u, a solution to (1.1) as the limit of the sequence u_k, the approximate solutions of (3.1). Each u_k has a finite time extinction which does not depend on the domain and is uniformly bounded by Proposition 4.2. Therefore, there exists a finite time extinction for u. $\square$

Remark 4.7. If we take an unbounded domain Ω instead of $\mathbb{R}^N$, we can use a sequence of approximate nested domains with finite measure in order to get the same conclusions. In particular, it has to be noted that the proofs of Propositions 4.1 and 4.2 are also valid for an unbounded domain with finite measure. $\square$

4.2. Nonextinction results. In this subsection we will show that for $\lambda > \lambda_{N,p}$ there is no finite time extinction for the solutions to (2.2) in bounded domains with positive initial data and, as a consequence, for the solutions of the Cauchy problem (1.1). After that, we will show that there exists at least a solution to (1.1) with $\mu_{N,p} < \lambda \le \lambda_{N,p}$ which does not have finite time extinction; the existence of the self-similar solution S for this range of λ plays a fundamental role in the proof of this result.

PROPOSITION 4.8. *Let $\Omega \subset \mathbb{R}^N$ a bounded domain and consider u, a solution of the untruncated problem* (2.2) *with $\lambda > \lambda_{N,p}$ and $1 < p < 2$. Then there exists no finite time extinction for u.*

Proof. In [12, sect. 8], it is shown, following the ideas in [11], that $w(x,t) = ct^{1/(2-p)}\phi_1(x)$ is a subsolution to the problem

$$\begin{cases} u_t - \Delta_p u &= \lambda \dfrac{u^{p-1}}{|x|^p}, & x \in \Omega,\, t > 0, \\ u(x,0) &= 0, & x \in \Omega, \\ u(x,t) &= 0, & x \in \partial\Omega,\, t > 0 \end{cases}$$

for $\lambda > \lambda_{N,p}$, where ϕ_1 verifies

$$\begin{cases} -\Delta_p \phi_1 &= \lambda_1(n) W_n(x) \phi_1^{p-1}, & x \in \Omega, \\ \phi_1 &= 0, & x \in \partial\Omega, \end{cases}$$

$\lambda_1(n)$ being the first eigenvalue of the previous problem, for n large enough such that $\lambda_1(n) < \lambda$. Then w is also a subsolution to the problem (2.2). Since w cannot have finite time extinction, there exists no finite time extinction for any solution to (2.2) (observe that $w(x,0) = 0$). □

Remark 4.9. As a consequence of Proposition 4.8, we conclude that there exists no finite time extinction for any solution to the corresponding Cauchy problem if $\lambda > \lambda_{N,p}$ and $u_0(x) > 0$ in a ball in $\mathbb{R}^N$. □

Moreover, if we take $\lambda > \mu_{N,p}$, we know that there exists a self-similar solution S to the Cauchy problem with zero initial data in the range $1 < p < 2N/(N+1)$ (see section 3.1). In this way we obtain for the same problem a solution with finite time extinction (the trivial solution) and a solution which does not have finite time extinction (S; observe that $S \notin L^s_{loc}(\mathbb{R}^N)$). We are showing that, in this range, there exists at least a solution of the Cauchy problem with positive initial data which does not have finite time extinction.

THEOREM 4.10. *If $1 < p < 2N/(N+1)$, $\mu_{N,p} < \lambda$, $u_0 \geq \delta > 0$ on a ball B and there exists a shift in time of S, $\overline{S}$, such that $u_0 \leq \overline{S}$, then there exists a solution of the Cauchy problem* (1.1) *with no finite time extinction.*

Proof. Assume that u_0 is bounded (if not, take $T_k(u_0)$). Consider the following iteration for $k > 0$:

$$\begin{cases} v_{kt} - \Delta_p v_k &= \lambda \dfrac{\tilde{v}_{k-1}^{p-1}}{|x|^p}, & x \in B_k,\, t > 0, \\ v_k(x,0) &= u_0(x), & x \in B_k, \\ v_k(x,t) &= S(x,t), & x \in \partial B_k,\, t > 0, \end{cases}$$

where $v_0 = \overline{S}$, $\tilde{v}_k = v_k$ in B_k, $\tilde{v}_k = S$ in $\mathbb{R}^N \setminus B_k$, B_k being the ball of radius k centered at the origin in $\mathbb{R}^N$. For a fixed $T > 0$, we make a separate study for the cases $1 < p < 2N/(N+2)$ and $2N/(N+2) \leq p < 2N/(N+1)$:

(a) Assume $1<p<2N/(N+2)$; in this range we know that $\overline{S} \in L^p(0,T;W^{1,p}(B_1))$ (see section 3.1). In addition, by means of a similar calculation to the one carried out in the proof of Lemma 2.3(a), we have

$$\frac{\overline{S}^{p-1}}{|x|^p} \in L^{p'}(0,T;W^{-1,p'}(B_1)).$$

Therefore, since $\overline{S}_t \in L^{p'}(0,T;W^{-1,p'}(B_1))$ too, and $v_1 = S$ on ∂B_1, we can apply the results in [18], obtaining the existence of a unique solution to the problem corresponding to $k = 1$, namely,

$$v_1 \in \mathcal{C}(0,T;L^2(B_1)) \cap L^p(0,T;W^{1,p}(B_1)).$$

Hence, $\tilde{v}_1 \in L^p(0,T;W^{1,p}(B_2))$ and $|x|^{-p}\tilde{v}_1^{p-1} \in L^{p'}(0,T;W^{-1,p'}(B_2))$, which in turn implies the existence of v_2, a solution to the problem corresponding to $k = 2$, etc. Moreover,

$$\begin{array}{rcll} v_{1t} - \Delta_p v_1 = \lambda \dfrac{\overline{S}^{p-1}}{|x|^p} & = & \overline{S}_t - \Delta_p \overline{S}, & x \in B_1,\ t > 0, \\ v_1(x,0) = u_0(x) & \leq & \overline{S}(x,0), & x \in B_1, \\ v_1(x,t) = S(x,t) & \leq & \overline{S}(x,t), & x \in \partial B_1,\ t > 0, \end{array}$$

and

$$\begin{array}{rcll} v_{1t} - \Delta_p v_1 = \lambda \dfrac{\overline{S}^{p-1}}{|x|^p} & \geq & \lambda \dfrac{S^{p-1}}{|x|^p} = S_t - \Delta_p S, & x \in B_1,\ t > 0, \\ v_1(x,0) = u_0(x) & \geq & 0 = S(x,0), & x \in B_1, \\ v_1(x,t) & = & S(x,t), & x \in \partial B_1,\ t > 0; \end{array}$$

therefore we conclude that

$$S \leq v_1 \leq v_0 = \overline{S} \qquad \text{in } B_1.$$

In a similar way, if we assume $S \leq v_{k-1} \leq \overline{S}$ in B_{k-1}, then v_k verifies $S \leq v_k \leq \overline{S}$ in B_k. If we now take a cut function φ and proceed as in the proof of Theorem 3.4, we can pass to the limit and obtain a solution v of the Cauchy problem with no finite time extinction, since the approximate solutions v_k imply that $v \geq S$.

(b) If $2N/(N+2) \leq p < 2N/(N+1)$, then $\overline{S} \in L^q_{loc}(\mathbb{R}^N)$ for $1 \leq q < s$ (see section 3.1). Assume $k = 1$ and take the approximate problems, for $n > 0$,

$$\left\{ \begin{array}{rcll} w_{nt} - \Delta_p w_n & = & \lambda f_n = \lambda T_n(|x|^{-p}\overline{S}^{p-1}), & x \in B_1,\ t > 0, \\ w_n(x,0) & = & u_0(x), & x \in B_1, \\ w_n(x,t) & = & T_n(S(x,t)), & x \in \partial B_1,\ t > 0. \end{array} \right.$$

Since the right-hand side in the previous equation and the boundary data are both bounded, we can apply the results in [18] to obtain the existence of a unique $w_n \in L^p(0,T;W^{1,p}(B_1))$. Moreover, we can show that $w_n \leq \overline{S}$, since $(w_n - \overline{S})_+ = 0$ on the parabolic boundary of the cylinder $B_1 \times (0,T)$,

$$(w_n - \overline{S})_t - (\Delta_p w_n - \Delta_p \overline{S}) = \lambda(T_n(|x|^{-p}\overline{S}^{p-1}) - |x|^{-p}\overline{S}^{p-1}) \leq 0,$$

and we can take $(w_n - \overline{S})_+$ as a test function to integrate the inequality above, which yields $w_n \leq \overline{S}$. In the same way we can show that

$$w_1 \leq w_2 \leq \cdots \leq w_n \leq \cdots \leq \overline{S}.$$

That is, we have a monotone sequence in $L^p(0,T;W^{1,p}(B_1))$ uniformly bounded by $\overline{S}$. If we define

$$v_1 = \lim_{n\to\infty} w_n \leq \overline{S},$$

we can proceed as in the proof of Theorem 3.4 (multiplying by $T_l(w_n\varphi^p)$ the equation satisfied by w_n, where φ is a cut function on a ball $B \subset B_1$ which does not depend on t) in order to show

$$\frac{1}{l}\int_0^T \int_{B\cap\{w_n<l\}} |\nabla w_n|^p \, dx \, dt \leq M,$$

where M does not depend on l.

Therefore there exists $v_1 \leq \overline{S}$ verifying the equation corresponding to $k = 1$ in $\mathcal{D}'(B \times (0,T))$, for any ball $B \subset B_1$ and $v_1(x,0) = u_0(x)$ in B_1, $v_1(x,t) = S(x,t)$ on ∂B_1, since v_1 is the pointwise limit of the approximate solutions w_n. By repeating this procedure for each $k > 1$, we obtain a solution v_k of the equation corresponding to k in $\mathcal{D}'(B_R \times (0,T))$, where B_R is any ball in $\mathbb{R}^N$, for k large enough.

In this way, we have found for $2N/(N+2) \leq p < 2N/(N+1)$ a sequence of functions uniformly bounded from above by $\overline{S}$ and, by recurrence, bounded from below by $S \leq v_k$ in B_k. Thus, using the same arguments as those in the proof of Theorem 3.4, we can find a solution v of the Cauchy problem (1.1) as the limit of v_k as $k \to \infty$, where $\mu_{N,p} < \lambda$ and $2N/(N+2) \leq p < 2N/(N+1)$, with no finite time extinction, since the lower bounds for the approximate solutions v_k imply that $v \geq S$. □

5. Further results.

Let $\Omega \subset \mathbb{R}^N$ be a bounded domain and take $1 < p < 2$ and $p < N$. Consider the problem

$$\begin{cases} u_t - \Delta_p u = \lambda V(x) u^{p-1}, \quad x \in \Omega \subset \mathbb{R}^N,\ t > 0,\ \lambda > 0, \\ u(x,0) = u_0(x) \in L^2(\Omega), \qquad u_0(x) \geq 0, \\ u(x,t) = 0,\ x \in \partial\Omega,\ t > 0, \end{cases} \tag{5.1}$$

with $V \in L^q(\Omega)$, $q > 1$, and $V(x) \geq 0$. One can show the existence of a solution to (5.1) for λ small enough and q large enough, respectively, by means of the techniques used in section 3. This is a particular case of the following general problem:

$$\begin{cases} u_t - \Delta_p u = \lambda V(x) f(u), \quad x \in \Omega \subset \mathbb{R}^N,\ t > 0, \\ u(x,0) = u_0(x) \in L^2(\Omega), \quad u_0(x) \geq 0, \\ u(x,t) = 0,\ x \in \partial\Omega,\ t > 0, \end{cases} \tag{5.2}$$

where we are assuming that $f(\sigma)$ is a continuous nondecreasing function for $\sigma \geq 0$, $f(0) = 0$, and $f(\sigma) > 0$ for $\sigma \in (0, M]$. In some cases, we can show the existence of a solution to (5.2) (the proofs of the following results only deal with the main estimates needed to pass to the limit in the way shown in section 3).

5.1. Global existence with unbounded initial data.

THEOREM 5.1.

(A) *If there exists some $\eta \in [0,2]$ such that $V \in L^{2/(2-\eta)}(\Omega)$ and*

$$\frac{f(\sigma)}{\sigma^{\eta-1}} \to C \geq 0 \qquad \text{as } \sigma \to \infty$$

then there exists a global solution to (5.2) *such that*

$$u \in L^\infty_{loc}([0,\infty); L^2(\Omega)) \cap L^p_{loc}((0,\infty); W_0^{1,p}(\Omega)).$$

(B) *If $1 < p < 2N/(N+2)$, ($s = N(2-p)/p > 2$), $V \in L^{s/(2-\eta)}(\Omega)$ for some $\eta \in [0,2]$ and*

$$\frac{f(\sigma)}{\sigma^{\eta-1}} \to C \geq 0 \qquad \text{as } \sigma \to \infty,$$

then there exists a global solution to (5.2) *verifying*

$$u \in L^\infty_{loc}([0,\infty); L^s(\Omega)) \qquad u^{(p+s-2)/p} \in L^p_{loc}((0,\infty); W_0^{1,p}(\Omega)).$$

Proof. (A) Let u be a solution to problem (5.2). For $T > 0$, we get, using Hölder and Young inequalities and integrating in $[0, T]$,

$$\frac{1}{2}\int_\Omega u^2(x,T)\,dx + \int_0^T \int_\Omega |\nabla u(x,t)|^p\,dx\,dt$$
$$\leq \frac{1}{2}\int_\Omega u_0^2(x)\,dx + C_q T\int_\Omega V^q(x)\,dx + \int_0^T \int_\Omega (f(u(x,t))u(x,t))^{q'}\,dx\,dt.$$

By hypothesis, there exists some $\eta \in [0,2]$ such that $V \in L^q(\Omega)$, with $q = 2/(2-\eta)$; this implies that

$$\frac{(f(\sigma)\sigma)^{q'}}{\sigma^2} = \left(\frac{f(\sigma)}{\sigma^{(q-2)/q}}\right)^{q'} = \left(\frac{f(\sigma)}{\sigma^{\eta-1}}\right)^{q'} \to C \quad \text{ as } \sigma \to \infty.$$

Hence $(f(u)u)^{q'} \leq Cu^2$; in other words,

$$\begin{aligned}\frac{1}{2}\int_\Omega u^2(x,T)\,dx \quad &+ \quad \int_0^T \int_\Omega |\nabla u(x,t)|^p\,dx\,dt \\ &\leq \quad \frac{1}{2}\int_\Omega u_0^2(x)\,dx + C_q T\int_\Omega V^q(x)\,dx + C\int_0^T \int_\Omega u^2(x,t)\,dx\,dt.\end{aligned}$$

Gronwall inequality implies that

$$\int_\Omega u^2(x,T)dx \leq e^{CT}\int_\Omega (u_0^2(x) + C_q T V^q(x))\,dx\,dt.$$

That is, $u(\cdot,T) \in L^2(\Omega)$, and so $u \in L^\infty_{loc}([0,\infty); L^2(\Omega))$. Moreover, we also get

$$\int_0^T \int_\Omega |\nabla u(x,t)|^p\,dx\,dt < \infty;$$

hence $u \in L^p_{loc}([0,\infty) : W_0^{1,p}(\Omega))$.

(B) By multiplying the equations satisfied by the solutions u_n of the truncated problems corresponding to (5.2) (they are bounded by Lemma 2.1) by u_n^{s-1} and integrating in $[0, T]$ for $T > 0$ we have, using Hölder and Young inequalities,

$$\frac{1}{s}\int_\Omega u_n^s(x,T)\,dx + \frac{\mu_{N,p}}{\lambda_{N,p}}\int_0^T \int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p\,dx\,dt$$
$$\leq \frac{1}{s}\int_\Omega u_0^s(x)\,dx + C_q T\int_\Omega V^q(x)\,dx + \int_0^T \int_\Omega (f(u(x,t))u_n^{s-1}(x,t))^{q'}\,dx\,dt.$$

By hypothesis, there exists some $\eta \in [0,2]$ such that $q = s/(2-\eta)$; this implies that

$$\frac{(f(\sigma)\sigma^{s-1})^{q'}}{\sigma^s} = \left(\frac{f(\sigma)}{\sigma^{(q-s)/q}}\right)^{q'} = \left(\frac{f(\sigma)}{\sigma^{\eta-1}}\right)^{q'} \to C \quad \text{ as } \sigma \to \infty.$$

Hence $(f(u_n)u_n^{s-1})^{q'} \leq Cu_n^s$; in other words,

$$\begin{aligned}\frac{1}{s}\int_\Omega u_n^s(x,T)\,dx \quad &+ \quad \frac{\mu_{N,p}}{\lambda_{N,p}}\int_0^T \int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p\,dx\,dt \\ &\leq \quad \frac{1}{s}\int_\Omega u_0^s(x)\,dx + C_q T\int_\Omega V^q(x)\,dx + C\int_0^T \int_\Omega u_n^s(x,t)\,dx\,dt.\end{aligned}$$

Gronwall inequality implies that

$$\int_\Omega u_n^s(x,T)\,dx \le e^{CT}\int_\Omega (u_0^s(x)+C_qTV^q(x))\,dx.$$

That is, passing to the limit in the nondecreasing sequence u_n, we get $u(x,T)\in L^s(\Omega)$, and so $u\in L^\infty_{loc}([0,\infty);L^s(\Omega))$. On the other hand, we also have $u^{(p+s-2)/p}\in L^p_{loc}((0,\infty);W_0^{1,p}(\Omega))$. □

Remark 5.2. If $1<p<2N/(N+2)$, then there exists η such that $p<\eta<2$ and $2/(2-\eta)<N/p$. Therefore, Theorem 5.1 implies the existence of a global solution to (5.2) for supercritical V and superdiffusive η. □

5.2. Some remarks about uniqueness.

LEMMA 5.3. *Consider the problem*

$$\left\{\begin{array}{rcll} u_t-\Delta_p u &=& \lambda f(u), & x\in\Omega\subset\mathbb{R}^N,\ t>0,\\ u(x,0) &=& 0, & \\ u(x,t) &=& 0, & x\in\partial\Omega,\ t>0. \end{array}\right. \tag{5.3}$$

If there exists some η such that $1<p<\eta\le 2$ and

$$\frac{f(\sigma)}{\sigma^{\eta-1}}\to C \qquad \text{as } \sigma\to 0,$$

then the unique solution $u\in C([0,T];L^s(\Omega))\cap L^p([0,T],W_0^{1,p}(\Omega))$ is $u\equiv 0$, where $s=2$ if $2N/(N+2)\le p<2$ and $s=N(2-p)/p$ if $1<p<2N/(N+2)$.

Proof. The case $2N/(N+2)\le p<2$ was essentially covered in [12, Lemma 8.6].

If $1<p<2N/(N+2)$, we follow a similar argument multiplying the equation by u^{s-1}. Then

$$\frac1s\int_\Omega u^s(x,T)dx+\frac{\mu_{N,p}}{\lambda_{N,p}}\int_0^T\int_\Omega|\nabla u^{(p+s-2)/p}(x,t)|^p\,dx\,dt$$
$$=\lambda\int_0^T\int_\Omega f(u(x,t))u^{s-1}(x,t)\,dx\,dt.$$

By Sobolev, we get

$$\frac1s\int_\Omega u^s(x,T)\,dx+C_{N,p}\int_0^T\left(\int_\Omega u^s(x,t)\,dx\right)^{(N-p)/N}dt$$
$$\le\lambda\int_0^T\int_\Omega f(u(x,t))u^{s-1}(x,t)\,dx\,dt.$$

Calling now $y(t)\equiv(\int_\Omega u^s(x,t)\,dx)^{1/s}$, for $T>0$ small enough, we obtain

$$0\le y^s(T)\le\int_0^T(C_{|\Omega|}\lambda y^{\eta+s-2}(t)-C_{N,p}y^{p+s-2}(t))\,dt.$$

Since $\eta>p$, this implies that $y(T)\equiv 0$ for every $T>0$, that is, $u\equiv 0$, and hence the limit verifies $u\equiv 0$. □

With respect to the nonuniqueness, assume f is concave in $[0,M]$ with

$$\int_0^M\frac{d\sigma}{f^{-1}(\sigma)}<\infty.$$

It may be checked that, in these hypotheses, $h(\sigma) = \sigma/f(\sigma)$ is an increasing function and $h(\sigma) \to 0$ as $\sigma \to 0$. In addition, if $\mu = \mu(t)$ is given by

$$\int_0^{\mu} \frac{d\sigma}{f(\sigma)} = t, \qquad t \geq 0,$$

then μ is well-defined, it is continuous, and nonnegative in $[0, t_0]$, where

$$t_0 = \int_0^M \frac{d\sigma}{f(\sigma)}$$

and μ solves (see [11])

$$\begin{cases} \dfrac{d\mu}{dt} &= f(\mu), \qquad 0 \leq t \leq t_0, \\ \mu(0) &= 0. \end{cases}$$

Take λ_1, the first eigenvalue of $-\Delta_p$, in the bounded domain Ω with zero boundary data and the corresponding normalized eigenfunction $\phi_1 \leq 1$ (see [11]). Then we can show the following lemma.

LEMMA 5.4. *Consider the problem* (5.2) *where* $1 < p < 2$, $V \equiv 1$, $u_0 \equiv 0$. *If* f *is concave in* $[0, M]$ *with*

$$\int_0^M \frac{d\sigma}{f^{-1}(\sigma)} < \infty$$

and

$$\frac{\sigma^{p-1}}{f(\sigma)} = \frac{h(\sigma)}{\sigma^{2-p}} \to C \qquad as\ \sigma \to 0,$$

where $C \in [0, \lambda_1^{-1})$, *then there exists* $\tau > 0$ *such that the solution of* (5.2) *is not unique in* $0 \leq t \leq \tau$, $0 \leq u \leq M$.

Remark 5.5. For instance, we can apply Lemmas 5.3 and 5.4 to the following particular cases:

1. If $f(\sigma) = \lambda\sigma^{\eta-1}$, where $p < \eta \leq 2$, we have uniqueness.
2. If $f(\sigma) = \lambda\sigma^{\eta-1}$, where $1 < \eta < p$ or $f(\sigma) = \lambda\sigma^{p-1}$ with $\lambda > \lambda_1$, then we have nonuniqueness. □

It has to be noted that nonuniqueness for zero initial data implies nonexistence of finite time extinction for the solutions of the corresponding problems with nonnegative initial data (see next subsection).

5.3. Finite time extinction. If u is a solution of (5.1), using the Hölder inequality and following the proof of Proposition 4.1, we get

$$||u(x,T)||_2 \leq ||u_0||_2 \left(1 - \frac{(2-p)\gamma|\Omega|^{1-p/2-p/N}T}{||u_0||_2^{2-p}}\right)_+^{1/(2-p)}$$

for $2N/(N+2) \leq p < 2$, $q > N/p$, $\lambda < (C||V||_q)^{-1}$ and $\gamma = \gamma(N, p, \Omega, \lambda) > 0$. For instance, in the particular case $V(x) = |x|^{-\beta}$, we obtain the result for $\beta < p$.

On the other hand, let $\mu_{N,p}$, s be defined as in section 3.1. If $1 < p < 2N/(N+2)$ ($s > 2$), then we can multiply the equation in the truncated problem corresponding to (5.1) (see section 2) by u_n^{s-1}, where u_n is the weak solution of the truncated problem. Then we get, using again the Hölder inequality and following the proof of Proposition 4.2, since the supersolution ϕ_n introduced in section 2 does not depend on the potential V,

$$||u_n(x,T)||_s \leq ||u_0||_s \left(1 - \frac{(2-p)\gamma T}{||u_0||_s^{2-p}}\right)_+^{1/(2-p)},$$

where $\lambda < C_{N,p}||V||_q^{-1}$ (the smaller $||V||_q$ is, the larger the range for λ again), and γ depends on N, p, λ and $||V||_q$ with $q = N/p$. Therefore, we get the result for the solution u to (5.1) obtained as the limit of the nondecreasing sequence u_n.

In this way we can see that there exists a finite extinction time for the solutions of (5.1) if λ is small enough.

Now consider ϕ_1, a solution to the eigenvalue elliptic problem

$$\left\{ \begin{array}{rcll} -\Delta_p \phi_1 & = & \lambda_1 V(x)\phi_1^{p-1}, & x \in \Omega \subset \mathbb{R}^N, \\ \phi & = & 0, & x \in \partial\Omega, \end{array} \right. \tag{5.4}$$

where $V \in L^q(\Omega)$, with $q > N/p$, and λ_1 is the first eigenvalue for $-\Delta_p$ with the weight V in Ω and zero boundary data, that is,

$$\lambda_1 = \inf_{\phi \in W_0^{1,p}(\Omega)} \frac{\int_\Omega |\nabla\phi|^p}{\int_\Omega V(x)\phi^p}.$$

Then, multiplying the equation in (5.1) by u, we get, for $\lambda < \lambda_1$,

$$\frac{1}{2}\frac{d}{dt}||u(x,t)||_2^2 + \gamma \int_\Omega |\nabla u(x,t)|^p \, dx \leq 0$$

with $\gamma > 0$, so that we get again the estimate above for the finite extinction time of u if $2N/(N+2) \leq p < 2$ (cf. proof of Proposition 4.1).

If we multiply the equation satisfied by the weak solution u_n of the truncated problem corresponding to (5.1) by u_n^{s-1}, and $\lambda << \lambda_1$, there exists $\gamma > 0$ such that

$$\frac{1}{s}\frac{d}{dt}||u_n(x,t)||_s^s + \gamma \int_\Omega |\nabla u_n^{(p+s-2)/p}(x,t)|^p \, dx \leq 0,$$

and we obtain again finite time extinction for the solutions of (5.1) obtained as the limit of the sequence u_n, $1 < p < 2N/(N+2)$, independently on the domain (cf. proof of Proposition 4.2).

In particular, if we are dealing with $V \in L^q(\mathbb{R}^N)$, $q > N/p$, and $\lambda < \lambda_1 \mu_{N,p}/\lambda_{N,p}$, then there exists a finite extinction time for the solutions of the corresponding Cauchy problem which are obtained as the limit as $k \to \infty$ of the solutions u_k of the Dirichlet problems (5.1) on $\Omega = B_k$, the ball of radius k centered at the origin in $\mathbb{R}^N$.

However, if $\lambda > \lambda_1$, then any solution to (5.1) cannot have finite time extinction, since $w(x,t) = ct^{1/(2-p)}\phi_1(x)$ is a positive subsolution of (5.1), for c small enough (see proof of Proposition 4.8). Hence there is a gap for the values of λ (the interval $[\lambda_1 \mu_{N,p}/\lambda_{N,p}, \lambda_1]$) in which the question of the existence of a finite extinction time remains open.

Therefore we can summarize the results obtained in the following statement.

PROPOSITION 5.6. *If $1 < p < 2$ and u is a solution of* (5.1) *then there exists a finite time $T^\star$ depending only upon N, p, λ, u_0 and $||V||_q$ such that*

$$u(\cdot, t) \equiv 0 \qquad \forall t \geq T^\star.$$

Moreover,

$$0 < T^\star \leq \gamma_1 ||u_0||_2^{2-p} |\Omega|^{\frac{p}{2} + \frac{p}{N-1}},\ 2N/(N+2) \leq p < 2,\ q > N/p,\ \lambda < \lambda_1,$$

$$0 < T^\star \leq \gamma_2 ||u_0||_s^{2-p},\ 1 < p < 2N/(N+2),\ q > N/p,\ \lambda < \lambda_1 \frac{\mu_{N,p}}{\lambda_{N,p}},$$

$$0 < T^\star \leq \gamma_2 ||u_0||_s^{2-p},\ 1 < p < 2N/(N+2),\ q = N/p,\ \lambda < \gamma_{N,p}^{-\frac{(N-p)}{N}} ||V||_q^{-1} \frac{\mu_{N,p}}{\lambda_{N,p}},$$

where γ_1 and γ_2 are positive constants depending only upon N, p, λ and $||V||_q$. Note that $|x|^{-p} \in \cap L^q_{loc}(\mathbb{R}^N)$ for every $q < N/p$.

Remark 5.7. A similar result is obtained for a solution of (5.2) where the general nonlinearity $f(\sigma)$ verifies $f(\sigma) \leq C\sigma^{p-1}$, for $1 < p < 2$, namely, there exists a finite extinction time if

1. $\lambda < \lambda_1/C$ for $2N/(N+2) \leq p \leq 2$.
2. $\lambda < \lambda_1 \mu_{N,p}/(C\lambda_{N,p})$ for $1 < p \leq 2N/(N+2)$. □

Acknowledgments. The authors want to thank their colleague Jesús García Azorero for many discussions and also the referee for his careful reading of the manuscript and very valuable suggestions.

REFERENCES

[1] P. Baras and J. Goldstein, *The heat equation with a singular potential*, Trans. Amer. Math. Soc., 294 (1984), pp. 121–139.

[2] P. Bénilan and M. G. Crandall, *The continuous dependence on φ of solutions of $u_t - \Delta\varphi(u) = 0$*, Indiana Univ. Math. J., 30 (1981), pp. 161–177.

[3] P. Bénilan, L. Boccardo, T. Gallouët, R. Gariepy, M. Pierre, and J. L. Vázquez, *An L^1-theory of existence and uniqueness of solutions of nonlinear elliptic equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 22 (1995), pp. 241–273.

[4] L. Boccardo and T. Gallouët, *Nonlinear elliptic equations with right hand side measures*, Comm. Partial Differential Equations, 17 (1992), pp. 641–655.

[5] L. Boccardo and F. Murat, *Almost everywhere convergence of the gradients of solutions to elliptic and parabolic equations*, Nonlinear Anal., 19 (1992), pp. 581–597.

[6] H. Brézis, *Analyse Fonctionnelle*, Masson, Paris, 1983.

[7] H. Brézis, T. Cazenave, Y. Martel, and A. Ramiandrisoa, *Blowup for $u_t - \Delta u = g(u)$ revisited*, Adv. Differential Equations, 1 (1996), pp. 73–90.

[8] E. DiBenedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[9] E. DiBenedetto and M. A. Herrero, *On the Cauchy problem and initial traces for a degenerate parabolic equation*, Trans. Amer. Math. Soc., 314 (1989), pp. 187–224.

[10] E. DiBenedetto and M. A. Herrero, *Non-negative solutions of the evolution p-Laplacian equation. Initial traces and Cauchy problem when $1 < p < 2$*, Arch. Rational Mech. Anal., 111 (1990), pp. 225–290.

[11] H. Fujita, *On Some Nonexistence and Nonuniqueness Theorems for Nonlinear Parabolic Equations*, Proc. Sympos. Pure Math. 18, AMS, Providence, RI, 1968.

[12] J. García Azorero and I. Peral Alonso, *Hardy inequalities and some critical elliptic and parabolic problems*, J. Differential Equations, 144 (1998), pp. 441–476.

[13] I. M. Gelfand, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Transl. Ser. 2, 29 (1963), pp. 295–381.

[14] G. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934.

[15] M. A. Herrero and J. L. Vázquez, *Asymptotic behaviour of the solutions of a strongly nonlinear parabolic problem*, Ann. Fac. Sci. Toulouse Math. (5), 3 (1981), pp. 113–127.

[16] D. D. Joseph and T. S. Lundgren, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Rational Mech. Anal., 49 (1973), pp. 241–269.

[17] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[18] J. L. Lions, *Quelques méthodes de resolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[19] F. Merle and L. A. Peletier, *Positive solutions of elliptic equations involving supercritical growth*, Proc. Roy. Soc. Edinburgh Sect. A, 118 (1991), pp. 49–62.

[20] I. Peral and J. L. Vazquez, *On the stability or instability of the singular solution of the semilinear heat equation with exponential term*, Arch. Rational Mech. Anal., 129 (1995), pp. 201–224.

[21] X. Xu, *On the initial-boundary-value problem for* $u_t - \operatorname{div}(|\nabla u|^{p-2}\nabla u) = 0$, Arch. Rational Mech. Anal., 127 (1994), pp. 319–335.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1295–1306

IDENTIFICATION OF TWO-PHASE FREE BOUNDARY ARISING IN PLASMA PHYSICS[*]

JUNE-YUB LEE[†] AND JIN KEUN SEO[‡]

Abstract. We try to estimate the shape and the location of two-phase free boundary which has been studied in [A. Friedman and Y. Liu, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4), 22 (1995), pp. 375–448] to model a stationary magnetohydrodynamics system. A sufficient condition is obtained to check whether a test disk is included in the plasma region D surrounded by a two-phase free boundary. In the test disk technique, only two simply verifiable conditions are used and indispensableness of the conditions is demonstrated using an example. The technique is applicable to select some of test disks placed in the domain Ω, which gives a rough guess on the shape of plasma region. Next we draw some geometrical properties of plasma region D when the domain Ω possesses a kind of convexity. It is proved that if Ω itself contains the mirror image of the right portion $\{x \in \Omega : x \cdot \xi > t\}$ of the domain with respect to a line $\{x : x \cdot \xi = t\}$ for all $t > t_0$, then so does the plasma region.

Key words. two-phase free boundary, test disk technique, dumbbell shaped domain, moving plane method, symmetric convex domain

AMS subject classifications. Primary, 35R35; Secondary, 31A25, 31B20, 76W05

PII. S0036141099351188

1. Introduction.

A two-phase free boundary problem is a mathematical model to find an interface between two disjoint domains on which solutions satisfy different types of governing equations. Such problems arise in various physical and engineering systems and an example we consider originates from a magnetohydrodynamics system which consists of vacuum and plasma region [6, 8]. During the last 20 years, significant progress in the study of free boundary problems has been made and many results regarding existence and regularity of the solutions have been obtained [2, 3, 4, 5]. However, information about global shape of the interfaces is more desired in many practical situations than regularity results.

Our main interest is to develop new techniques to estimate the size, the location, and some geometric properties of the region D surrounded by a two-phase free boundary. In particular, we consider a free boundary problem in a toroidally symmetric tokamak machine with two-dimensional cross section Ω. Not much has been known in this direction and many fundamental questions are still yet to be answered. For example, we still don't know whether D is convex provided Ω is convex. Though this particular question in the case of one-phase free boundary has been studied in several papers [1, 7, 10], the arguments cannot be directly applicable to our two-phase problem. Identifying the exact shape of the free boundary is not an easy task and it is partially due to the global dependency on the geometry of Ω and counterexamples of uniqueness of the plasma region.

We start with mathematical description of our free boundary problem and readers interested in the derivation of this problem and its physical meaning may consult our

[*] Received by the editors February 4, 1999; accepted for publication (in revised form) November 9, 1999; published electronically May 15, 2000. This work was supported in part by Korea Science and Engineering Foundation (KOSEF) under 1999-Basic Research Grant Program and in part by Korea Research Foundation (KRF) under 1998-BSRI support program.

http://www.siam.org/journals/sima/31-6/35118.html

[†] Department of Mathematics, Ewha Woman's University, Seoul 120-750, Korea (jylee@math.ewha.ac.kr).

[‡] Department of Mathematics, Yonsei University, Seoul 120-749, Korea (seoj@bubble.yonsei.ac.kr).

previous paper [8], the paper by Friedman and Liu [6], and its references. Suppose Ω is a bounded domain in $\mathbf{R}^n$ with $\mathcal{C}^2$ boundary $\partial\Omega$ and c, μ are given positive and nonnegative constants, respectively. In [6], Friedman and Liu considered a free boundary problem to find a solution u, a positive constant λ, and an interface between the plasma region D and the vacuum region satisfying the following equations:

$$\Delta u = 0 \text{ in } \{x : u(x) > 0\} \ \text{ with } \ u|_{\partial\Omega} = c, \tag{1.1}$$

$$\Delta u + \lambda u = 0 \text{ in } \{x : u(x) \leq 0\} \ \text{ with } \int_{\{u\leq 0\}} u^2 = 1, \tag{1.2}$$

$$\left|\nabla u^+\right|^2 - \left|\nabla u^-\right|^2 = \mu^2 \text{ along the interface } \partial D, \tag{1.3}$$

where the plasma region D is the interior of the set $\{x \in \Omega : u(x) \leq 0\}$ and where ∇u^- and ∇u^+ denote the nontangential limits of ∇u from D and from the vacuum region $\Omega\backslash D$, respectively. It is proved that there exists a minimizer u_μ of the following minimizing problem:

$$\mathcal{M}_\mu^\Omega \left| \begin{array}{l} \text{Minimize} \quad J_\mu(u) := \int_\Omega |\nabla u|^2 dx - \mu^2 \left|\{x \in \Omega : u(x) \leq 0\}\right| \\ \text{within the class} \quad \mathcal{K} = \{u \in H^1(\Omega) \ : \ u|_{\partial\Omega} = c, \ \int_{\{u\leq 0\}} u^2 = 1\}, \end{array}\right.$$

where $|D|$ denotes Lebesgue measure of set D. It is also proved that the minimizer u_μ^Ω satisfies (1.1)–(1.3) and for $n = 2$ the boundary of the plasma region ∂D_μ^Ω is smooth. Throughout this paper, u_μ^Ω denotes a minimizer of $\mathcal{M}_\mu^\Omega$ and $D_\mu^\Omega = \mathit{interior}\{u_\mu^\Omega \leq 0\}$ denotes the corresponding plasma region. The subscript$_\mu$ or the superscript$^\Omega$ will be omitted when there is not confusion. Note that the plasma region D_μ^Ω is not a single valued function with respect to μ and Ω but depends on the choice of u_μ^Ω since there might exist many minimizers u_μ for given μ and Ω.

We present in Lemma 2.1 that $|D_\mu|$ increases as μ increases, regardless of the choices of u_μ, and the energy functional J_μ is differentiable almost everywhere as shown below,

$$J_\mu(u_\mu) := J_0(u_0) - \int_0^\mu 2\lambda |D_\lambda| d\lambda.$$

In order to investigate the location and the size of D_μ, it is natural to check if a test ball in the domain is included in D_μ. Our main result in Theorem 2.2 is about a sufficient condition for a ball B to be included in the plasma region D_μ. We state the theorem in two-dimensional case although the arguments could be easily extended on higher dimension provided that the free boundary is sufficiently smooth. Let B be an open disk contained in Ω of size large enough to satisfy $\frac{|\partial B|}{|B|} \leq \mu$. Suppose the solution h_B of the Dirichlet–Laplace problem, $\Delta h_B = 0$ in $\Omega \setminus B$ with the boundary data $h_B = 1$ on $\partial\Omega$ and $h_B = 0$ on ∂B, satisfies a testing condition

$$|\nabla h_B| \leq \mu \text{ on } \partial B;$$

then the test disk B is included in the plasma region D_μ

$$B \subset D_\mu.$$

Two brief comments can be made regarding the theorem. First, the theorem contains useful tools to guess the plasma region by placing many test disks on the domain and selecting some of them. Second, the size limit condition $\frac{|\partial B|}{|B|} \leq \mu$ in the theorem is indispensable and a theorem conjecture without this condition has a counterexample which is shown in Example 2.3.

In section 3, we investigate some geometrical properties of plasma region D^{Ω}. This kind of work is possible when there is some limitation on the domain Ω and two natural domain properties are maybe symmetry and convexity. Our previous paper [8] gives some results on symmetric convex domains. In order to improve such results, we introduce a new concept which is named mirror covering domain. A domain Ω is called a mirror covering domain with respect to a line $T_\xi(t_0)$ if Ω itself contains the mirror image of the right portion $\{x \in \Omega : x \cdot \xi > t\}$ of the domain with respect to a line $\{x : x \cdot \xi = t\}$ for all $t > t_0$. In Theorem 3.1, we obtain mirror image covering properties which says that if the domain Ω is a mirror covering domain with respect to $T_\xi(t_0)$, then so does the plasma region. Our previous result in [8] proves a similar theorem, that is, if Ω is symmetric and convex with respect to x_2-axis, then so is D_μ. Our new theorem significantly improves our old result and is also applicable to more general domains which need not to be symmetric. The proof of the theorem is based on the moving plane method which was used in the paper by J. Serrin [11] who deals with one-phase free boundary.

2. The size and the location of the plasma region. The problem $\mathcal{M}_\mu^\Omega$ might have more than one solution (see [8]) and, in such a case, u_μ denotes any possible minimizers and D_μ denotes corresponding plasma region.

LEMMA 2.1. *For $\mu_1 < \mu_2$, $J(\mu) := J_\mu(u_\mu)$ and the corresponding energy $\int_\Omega |\nabla u_\mu|^2$ satisfies the following inequalities:*

$$(\mu_2^2 - \mu_1^2)|D_{\mu_1}| < J(\mu_1) - J(\mu_2) < (\mu_2^2 - \mu_1^2)|D_{\mu_2}|, \tag{2.1}$$

$$\int_\Omega |\nabla u_{\mu_1}|^2 < \int_\Omega |\nabla u_{\mu_2}|^2, \tag{2.2}$$

and the energy functional J is uniquely characterized by

$$J(\mu) = J(0) - \int_0^\mu 2\lambda |D_\lambda| d\lambda. \tag{2.3}$$

Proof. It is obvious that $J_{\mu_2}(u_{\mu_1}) \neq J_{\mu_2}(u_{\mu_2})$. Suppose not; u_{μ_1} is also a minimizer of the problem $\mathcal{M}_{\mu_2}^\Omega$ and therefore from (1.3), u_{μ_1} has to satisfy

$$|\nabla u_{\mu_1}^+|^2 - |\nabla u_{\mu_1}^-|^2 = \mu_2^2 \quad \text{on } \partial D_{\mu_1},$$

which is a contradiction since a minimizer u_{μ_1} for $\mathcal{M}_{\mu_1}^\Omega$ has μ_1^2 gradient square jump. Similarly, $J_{\mu_1}(u_{\mu_1}) \neq J_{\mu_1}(u_{\mu_2})$. Since u_{μ_1} and u_{μ_2} are minimizers of $\mathcal{M}_{\mu_1}^\Omega$ and $\mathcal{M}_{\mu_2}^\Omega$, respectively,

$$J_{\mu_1}(u_{\mu_1}) < J_{\mu_1}(u_{\mu_2}) = J_{\mu_2}(u_{\mu_2}) + (\mu_2^2 - \mu_1^2)|D_{\mu_2}|, \tag{2.4}$$

$$J_{\mu_2}(u_{\mu_2}) < J_{\mu_2}(u_{\mu_1}) = J_{\mu_1}(u_{\mu_1}) - (\mu_2^2 - \mu_1^2)|D_{\mu_1}|. \tag{2.5}$$

These inequalities give the lower and the upper bounds of $J(\mu_1) - J(\mu_2)$ in (2.1) which states that $J(\mu)$ is monotone decreasing and Lipschitz continuous and D_μ is increasing with respect to μ. (Note: $|D_\mu|$ could depend on the choice of a minimizer.)

To prove (2.2), we rewrite (2.4) in terms of $\int_\Omega |\nabla u_{\mu_1}|^2$ and $\int_\Omega |\nabla u_{\mu_2}|^2$.

$$\int_\Omega |\nabla u_{\mu_1}|^2 - \mu_1^2 |D_{\mu_1}| < \int_\Omega |\nabla u_{\mu_2}|^2 - \mu_1^2 |D_{\mu_2}| \tag{2.6}$$

and $|D_{\mu_1}| < |D_{\mu_2}|$ proves (2.2).

Since $|D_\mu|$ is monotone increasing and $\lim_{\mu\to\infty} |D_\mu| = \Omega$, $|D_\mu|$ as a function of μ is continuous except on countably many points. Therefore, (2.1) proves that $J(\mu)$ is differentiable almost everywhere and $J(\mu) = J(0) - \int_0^\mu 2\lambda |D_\lambda| d\lambda$. □

We assume $c = 1$ and $n = 2$ for simplicity in the later part of the section. For a given open subset F of Ω with smooth boundary, let h_F be the solution of the Dirichlet problem

$$\begin{aligned} \Delta h_F = 0 \quad &\text{in} \quad \Omega \setminus \bar{F}, \\ h_F|_F \equiv 0\ , \quad &h_F|_{\partial\Omega} = 1, \end{aligned}$$

and $\lambda(F)$ be the smallest eigenvalue of Δ for the Dirichlet problem in the domain F.

THEOREM 2.2. *Let B be an open disk contained in Ω with $\frac{|\partial B|}{|B|} \le \mu$. Suppose that the harmonic function h_B satisfies the estimate*

$$|\nabla h_B^+| \le \mu \text{ on } \partial B,$$

where ∇h_B^+ denotes the gradient of h_B from the outside of B. Then

$$B \subset D_\mu.$$

Proof. For simplicity of the notation, let $u := u_\mu$ and $D := D_\mu$. To derive a contradiction, assume $D^* := D \cup B \neq D$. Let $\tilde{u}$ be a function defined as the normalized first eigenfunction of Δ with Dirichlet boundary condition in D^* and the harmonic function h_{D^*} in $\Omega \setminus \overline{D}^*$; then it satisfies

$$J_\mu(\tilde{u}) = J_\mu(h_{D^*}) + \lambda(D^*). \tag{2.7}$$

Therefore, the fact that u is a minimizer of J_μ leads

$$J_\mu(\tilde{u}) - J_\mu(u) = J_\mu(h_{D^*}) + \lambda(D^*) - J_\mu(h_D) - \lambda(D) \ge 0.$$

Since $\lambda(D^*) < \lambda(D)$,

$$J_\mu(h_{D^*}) - J_\mu(h_D) > 0. \tag{2.8}$$

On the other hand, it is easy to derive the following inequalities from the assumption $|\nabla h_B^+| \le \mu$ on ∂B using the maximum principle and the Hopf lemma,

$$|\nabla h_{D^*}^+| < \mu \quad \text{on } \partial B \cap \partial D^* \quad \text{and} \quad |\nabla h_{D^*}^+| < |\nabla u^+| \quad \text{on } \partial D \cap \partial D^*. \tag{2.9}$$

Using integration by parts over the region where u is harmonic,

$$\int_{\Omega\setminus\bar{D}} |\nabla u|^2 = \int_{\partial\Omega} \frac{\partial u}{\partial \nu} = \int_{\partial D} |\nabla u^+|, \tag{2.10}$$

where ν denotes the unit out normal vector to the boundary. Similarly, we obtain

$$\int_{\Omega\setminus\bar{D}^*} |\nabla h_{D^*}|^2 = \int_{\partial D^*} |\nabla h_{D^*}^+|. \tag{2.11}$$

Using (2.9), (2.10), and (2.11),

$$\begin{aligned}(2.12)\ & J_\mu(h_{D^*}) - J_\mu(h_D)\\ &= \int_{\partial D^*} |\nabla h_{D^*}^+| - \int_{\partial D} |\nabla u^+| + \mu^2|D| - \mu^2|D^*|\\ &= \int_{\partial D\cap\partial D^*} \left(|\nabla h_{D^*}^+| - |\nabla u^+|\right) + \int_{\partial B\setminus\bar{D}} |\nabla h_{D^*}^+| - \int_{\partial D\cap B} |\nabla u^+| - \mu^2|E|\\ &< \int_{\partial D\cap\partial D^*} \left(|\nabla h_{D^*}^+| - |\nabla u^+|\right) + \mu\left(|\partial B\setminus\bar{D}| - |\partial D\cap B| - \mu|E|\right)\\ &\le \mu\left(|\partial B\setminus\bar{D}| - |\partial D\cap B| - \mu|E|\right),\end{aligned}$$

where $E = B\setminus D \neq \emptyset$.

In order to derive a contradiction using (2.8) and (2.12), it suffices to prove the following inequality:

$$I := |\partial B\setminus\bar{D}| - |\partial D\cap B| - \mu|E| \le 0.$$

This quantity is purely geometric and we can, without loss of generality, assume that the arc $\Gamma = \partial B\setminus\bar{D}$ has only one connected component since we can estimate total I by adding the values of I for each components in the case of multicomponent Γ. For simplicity, we assume that B is centered at the origin with radius ρ and the arc Γ is in the range $\theta = 0$ and $\theta = \alpha$. Let $L(t,\theta)$ be the ray joining $(t\cos\theta, t\sin\theta)$ to $(\rho\cos\theta, \rho\sin\theta)$ and $r(\theta)$ be the smallest nonnegative number such that $L(t,\theta)$ does not intersect D for all $r(\theta) < t < \rho$. It is easy to see that the $(r(\theta)\cos\theta, r(\theta)\sin\theta)$ lies on the set $\partial D\cap B$ and

$$\begin{aligned}|\partial D\cap B| &\ge \int_0^\alpha r(\theta)d\theta,\\ |E| &\ge \int_0^\alpha \frac{1}{2}\left[\rho^2 - r^2(\theta)\right]d\theta.\end{aligned}$$

Therefore,

$$\begin{aligned}I &\le \alpha\rho - \int_0^\alpha r(\theta)d\theta - \mu\int_0^\alpha \frac{1}{2}\left[\rho^2 - r^2(\theta)\right]d\theta\\ &= \int_0^\alpha \left[\rho - r(\theta)\right]\left[1 - \mu\frac{\rho + r(\theta)}{2}\right]d\theta\\ &\le 0.\end{aligned}$$

The last inequality is true since the assumption $\frac{|\partial B|}{|B|} = \frac{2}{\rho} \le \mu$ implies $\frac{1}{2}\mu\rho \ge 1$. This completes the proof. □

In Theorem 2.2, the condition $\frac{|\partial B|}{|B|} \le \mu$ is quite unusual and one might think this condition should be removed. We will, however, show that the condition is indispensable in the theorem by constructing a disk B such that $B\cap D_\mu = \emptyset$ even though $|\nabla h_B| \le \mu$ on ∂B for given $\Omega, \mu = 1$. (μ is set to be 1 for simplicity of the description.)

Let u^Ω denote a minimizer of the problem $\mathcal{M}_\mu^\Omega$ as usual and let $\tilde{u}^\Omega$ denote a minimizer of the functional

$$(2.13)\qquad \tilde{J}^\Omega(\phi) = \int_\Omega |\nabla\phi|^2 - \mu^2|\{\phi = 0\}|$$

within the class $\tilde{\mathcal{K}} = \{\phi \in H^1(\Omega) : \phi = 1 \text{ on } \partial\Omega\}$. This new problem is identical to $\mathcal{M}_\mu^\Omega$ except that $\int_{\{u\leq 0\}} u^2 = 1$ condition is missing. Note that the solutions of these problems are known when Ω is a disk B_a of radius a. (Detailed computation can be found in [8].) In summary, the minimizers u^{B_a} and $\tilde{u}^{B_a}$ are positive outside of a disk of radius r_c and the corresponding energy functional values can be explicitly computed as follows:

$$J_{\mu=1}^{B_a}(u^{B_a}) = \inf_{0<r<a} \left(\frac{2\pi}{\log a/r} + \frac{\lambda(B_1)}{r^2} - \pi r^2 \right)$$

and

$$\tilde{J}_{\mu=1}^{B_a}(\tilde{u}^{B_a}) = \inf_{0<r<a} \left(\frac{2\pi}{\log a/r} - \pi r^2 \right).$$

For example, $J_1^{B_e} \approx 6.045$ with $r_c \approx 1.550$, $\tilde{J}_1^{B_e} = 0$ with $\tilde{u}^{B_e} \equiv 1$ for $a = e \approx 2.718$ and $J_1^{B_{2e}} \approx -30.989$ with $r_c \approx 4.315$, $\tilde{J}_1^{B_{2e}} \approx -31.300$ with $r_c \approx 4.311$ for $a = 2e \approx 5.437$.

Example 2.3. *Let Ω be a dumbbell shaped domain consists of two disks and a narrow connecting bridge:*

$$\Omega = B_L \cup B_R \cup T_\epsilon, \tag{2.14}$$

where $B_L = B_e(0,0)$ is a disk centered at the origin and of radius e, $B_R = B_{2e}(4e,0)$ of radius $2e$, and T_ϵ is a narrow bridge $\{(x_1,x_2) : e-\epsilon < x_1 < 2e+\epsilon, |x_2| < \epsilon\}$. If ϵ is sufficiently small, a test disk $B = B_1(0,0)$ satisfies

$$|\nabla h_B| \leq \mu = 1. \tag{2.15}$$

However, the intersection of the plasma region D_μ and the test disk B is empty

$$D_\mu \cap B = \emptyset. \tag{2.16}$$

Proof. It is easy to show that the gradient of h_B on ∂B is bounded by $\mu = 1$. Define $w(x) = \log|x|$ in $1 \leq |x| \leq e$. Then $\Delta w = 0$ in $B_L \backslash B$ and $w|_{\partial B_L} = 1, w|_{\partial B} = 0$. From the maximum principle, $h_B \leq w$ in $B_L \setminus \bar{B}$, therefore, $|\nabla h_B(x)| \leq |\nabla w(x)| = 1$ on $|x| = 1$ by the Hopf lemma. Hence the disk B satisfies the condition $|\nabla h_B| \leq \mu = 1$.

Next we want to show that B does not intersect with D_μ. Note that this may happen since one of two conditions in Theorem 2.2 is missing; $\frac{|\partial B|}{|B|} = 2 \not\leq \mu = 1$. For sufficiently small ϵ, it is possible to prove that

$$D_\mu \subseteq B_{e-\epsilon}(0,0) \cup B_{2e-\epsilon}(4e,0). \tag{2.17}$$

Here we will just give a brief sketch of the proof of (2.17). Suppose that there exists a point $p \in \partial D_\mu$ in the ϵ-neighborhood of $\partial\Omega$, *distance*$(p, \partial\Omega) \leq \epsilon$; then $|\nabla u_\mu^+(p)| > \frac{C}{\epsilon}$ for some fixed constant C and $|\nabla u_\mu^-(p)| > \frac{C}{\epsilon} - \mu^2$ from the interface condition (1.3). Therefore, for $\epsilon \leq \frac{C}{\mu^2}$, p must lie on the boundary of the negative part of the plasma region, $p \in \partial\{u_\mu < 0\}$. Recall that D_μ has exactly one connected negative set and the negative set has finite measure bounded below since the smallest eigenvalue of the Laplacian operator on the negative set is bounded, $\lambda(\{u_\mu < 0\}) < J^\Omega(\mu) + \mu^2|\Omega|$. So there exists a nonzero measure connected negative set near the ϵ-neighborhood of the boundary $\partial\Omega$. It draws a contradiction to the fact that $J^\Omega(\mu)$ is bounded since

$|\nabla u^+|$ is of order $\frac{1}{\epsilon}$ along the boundary of the negative set with nonzero length and the harmonic function defined on $\Omega \setminus \bar{D}$ generates unbounded energy near the point p as ϵ approaches 0.

Hence the plasma region D_μ is away from ϵ-neighbor of $\partial\Omega$, that is, $D_\mu = D_L \cup D_R$ where $D_L = D_\mu \cap B_L$ and $D_R = D_\mu \cap B_R$. Also, $u_\mu < 0$ on either D_L or D_R and $u_\mu \equiv 0$ in the other set, if it is not empty. Let u_μ^ϵ be defined as follows: $u_\mu^\epsilon = u_\mu$ in D_μ, $u_\mu^\epsilon = 1$ in $\Omega \setminus B_R \cup B_L$, and u_μ^ϵ is the harmonic function in $B_R \cup B_L \setminus D_\mu$ with boundary data $u_\mu^\epsilon = 1$ on $\partial(B_L \cup B_R)$ and $u_\mu^\epsilon = 0$ on ∂D_μ. Then the energy difference is quite small:

$$J_\mu^\Omega(u_\mu^\epsilon) = J_\mu^\Omega(u_\mu) + O(\epsilon) \tag{2.18}$$

and we can view $J_\mu^\Omega(u_\mu^\epsilon)$ as a sum of contributions from B_L and from B_R, separately. Two possible cases exist. First, B_L contains $u_\mu < 0$ set and B_R does not. Second, B_R does and B_L does not.

Case 1. $J_\mu^\Omega(u_\mu^\epsilon) \geq J_\mu^{B_L}(u_\mu^{B_L}) + \tilde{J}_\mu^{B_R}(\tilde{u}_\mu^{B_R}) \approx 6.045 - 31.300 = -25.255.$

Case 2. $J_\mu^\Omega(u_\mu^\epsilon) \geq \tilde{J}_\mu^{B_L}(\tilde{u}_\mu^{B_L}) + J_\mu^{B_R}(u_\mu^{B_R}) \approx 0.000 - 30.989 = -30.989.$

Since $J_\mu^\Omega(u_\mu^\epsilon) \geq J_\mu^\Omega(u_\mu) = J_\mu^\Omega(u_\mu^\epsilon) - O(\epsilon)$, we can conclude that the second case gives the minimal energy. Thus, $\{u_\mu < 0\} \subset B_R$ and

$$\tilde{J}_\mu^{B_L}(u_\mu) \leq O(\epsilon). \tag{2.19}$$

Now we want to prove $D_L = \emptyset$. At a glance over (2.19), one may guess that it must be $u_\mu \approx 1$ in B_L; however, there exists a counterexample of such a conclusion. To avoid such a mistake, the radius of B_L and μ should be taken into account. Suppose D_L is not an empty set; then $u_\mu \equiv 0$ in D_L and the interface condition (1.3) gives $|\nabla u_\mu| = \mu = 1$ on ∂D_L. Since $u_\mu = 1 + O(\epsilon)$ on ∂B_L, we have

$$\begin{aligned}
\tilde{J}_\mu^{B_L}(u_\mu) &= \int_{B_L} |\nabla u_\mu|^2 - |D_L| \\
&= \int_{\partial B_L} \frac{\partial u_\mu}{\partial \nu} u_\mu - |D_L| \\
&= \int_{\partial B_L} \frac{\partial u_\mu}{\partial \nu} - |D_L| + O(\epsilon) \\
&= \int_{\partial D_L} |\nabla u_\mu| - |D_L| + O(\epsilon) \\
&= |\partial D_L| - |D_L| + O(\epsilon).
\end{aligned}$$

Therefore it follows from (2.19) that

$$|\partial D_L| - |D_L| \leq O(\epsilon),$$

that is,

$$\frac{|\partial D_L|}{|D_L|} \leq 1 + O(\epsilon).$$

It follows from an elementary geometry that $|D_L| \geq 4\pi + O(\epsilon)$ to satisfy the above perimeter to area ratio. Let $r_0 = \sup\{|x| : x \in D_L\}$ and x_0 be a point on ∂D_L such that $|x_0| = r_0$. Then it must be $r_0 \geq 2 + O(\epsilon)$. Let H be the harmonic function in $B_L \setminus \bar{B}_{r_0}$ with the boundary condition $H = u_\mu$ on ∂B_L and $H = 0$ on ∂B_{r_0}. By maximum principle,

$$\begin{aligned} 1 = \mu = |\nabla u_\mu(x_0)| \geq |\nabla H(x_0)| &= \frac{1}{r_0(1-\log r_0)} + O(\epsilon) \\ &\geq \frac{1}{2(1-\log 2)} + O(\epsilon) > 1.629 + O(\epsilon), \end{aligned}$$

which is not possible. This proves $D_L = \emptyset$. □

This example shows that the condition $\frac{|\partial B|}{|B|} \leq \mu$ is indispensable in the disk covering theorem. D_μ can be exactly found if the disk covering technique method is able to check whether a point (or a disk with arbitrary small radius) is included inside of D_μ; however, this is not possible. As μ gets larger, the radius of the test disk can be chosen smaller, so the technique allows us to find a better approximation of the shape of the free boundary. It seems reasonable that detection of the interface with large jump of the normal derivatives along the interface is easier than that with small jump corresponding to small μ.

3. The mirror image covering properties.

In this section we want to find geometric properties of the plasma region D in some class of domains Ω. Our basic motivation is derived from the idea of the moving plane method used by Serrin [11]. We now introduce a new concept of mirror covering domain to describe our results. For a real number $t \in \mathbf{R}$ and a unit vector $\xi \in \mathbf{R}^2$, let us denote the hyperplane with normal vector ξ passing through $t\xi$ by $T_\xi(t) = \{x : x \cdot \xi = t\}$, the right-hand side portion of the domain by $\Sigma_\xi^\Omega(a) = \{x \in \Omega : x\cdot\xi > a\} = \Omega \cap \cup_{t>a} T_\xi(t)$, and the reflected image of $\Sigma_\xi^\Omega(a)$ with respect to $T_\xi(a)$ by $\tilde{\Sigma}_\xi^\Omega(a) = \{x' : x' = x + 2(\xi\cdot x - a)\xi, x \in \Sigma_\xi^\Omega(a)\}$. Then a domain Ω is called a mirror covering domain with respect to a line $T_\xi(t_0)$ if $\tilde{\Sigma}_\xi^\Omega(t) \subset \Omega$ for all $t > t_0$.

THEOREM 3.1. *For given $\mu \geq 0$, let u be a minimizer of $\mathcal{M}_\mu^\Omega$ and D be the corresponding plasma region. Suppose Ω is a mirror covering with respect to $T_\xi(a)$,*

$$\tilde{\Sigma}_\xi^\Omega(t) \subset \Omega \text{ for all } t > a. \tag{3.1}$$

Then so is the plasma region D,

$$\tilde{\Sigma}_\xi^D(a) \subset D. \tag{3.2}$$

Proof. Let

$$t_0 := \inf\{t \geq a : \tilde{\Sigma}_\xi^D(t) \subset D\}.$$

It suffices to prove $t_0 = a$. Suppose not, that is, $t_0 > a$. Then, as in the proof of Serrin [11], the following two events may occur: (i) $\tilde{\Sigma}_\xi^D(t_0)$ becomes internally tangent to the boundary of D at some point P not on $T_\xi(t_0)$. (ii) $T_\xi(t_0)$ is orthogonal to the boundary of D.

Let us introduce the reflected function v defined as follows:

$$v(x') := u(x) \quad \text{for } x' \in \tilde{\Sigma}_\xi^\Omega(t_0), \tag{3.3}$$

where x' is the reflected point of x across $T_\xi(t_0)$. Let $w := v - u$ in $\tilde{\Sigma}_\xi^\Omega(t_0)$. Then w satisfies

$$\begin{aligned}(\Delta+\lambda)w &= 0 \quad \text{in } \tilde{\Sigma}_\xi^D(t_0),\\ \Delta w &= 0 \quad \text{in } \tilde{\Sigma}_\xi^\Omega(t_0)\setminus \bar{D},\\ w &\geq 0 \quad \text{on } \partial\tilde{\Sigma}_\xi^D(t_0) \text{ and } \partial\tilde{\Sigma}_\xi^\Omega(t_0).\end{aligned}$$

Hence, using the monotonicity property of the first eigenvalue λ of the domain D and the maximum principle,

$$w > 0 \text{ in } \tilde{\Sigma}_\xi^D(t_0), \tag{3.4}$$

$$w > 0 \text{ in } \tilde{\Sigma}_\xi^\Omega(t_0)\setminus \bar{D}. \tag{3.5}$$

Using these inequality properties, we want to draw a contradiction to case (i) and case (ii), respectively. The proof is rather technical and the proof for the second case requires quite tedious computation.

Case (i). Since $w = 0$ at the contact point $P \in \partial\tilde{\Sigma}_\xi^D(t_0)$, by the Hopf lemma

$$\nabla w^+(P)\cdot\nu(P) > 0 \quad \text{and} \quad \nabla w^-(P)\cdot\nu(P) < 0,$$

where ∇w^+ and ∇w^- denote the gradients of w from outside and inside of $\tilde{\Sigma}_\xi^D(t_0)$, respectively, and $\nu(P)$ denotes the outer normal vector of $\partial\tilde{\Sigma}_\xi^D(t_0)$. Therefore,

$$|\nabla v^+(P)| = \nabla v^+(P)\cdot\nu(P) > \nabla u^+(P)\cdot\nu(P) = |\nabla u^+(P)|$$

and similarly,

$$|\nabla v^-(P)| < |\nabla u^-(P)|.$$

Hence by the inequalities above and the transmission condition (1.3) on the free boundary, we obtain a contradiction,

$$\mu^2 = |\nabla v^+(P)|^2 - |\nabla v^-(P)|^2 > |\nabla u^+(P)|^2 - |\nabla u^-(P)|^2 = \mu^2.$$

Case (ii). Let P be an orthogonal intersection point of $T_\xi(t_0)$ and D. We may assume $\nu(P) = e_2$ without loss of generality. Let $X_{\pm h} := P \pm he_1 + he_2$ and $Y_{\pm h} := P \pm he_1 - he_2$. For sufficiently small $h > 0$, $X_{\pm h}$ are in the vacuum region and $Y_{\pm h}$ are in the plasma region since the free boundary in two dimensions is smooth. Thus, we can derive two Taylor expansions for u near $X_{\pm h}$ and $Y_{\pm h}$, separately. Using the fact that u^+ is harmonic in $\Omega\setminus\overline{D}$,

$$\begin{aligned}u(X_{\pm h}) &= \partial_2 u^+(P)h \pm \partial_{12}u^+(P)h^2 + \frac{1}{2}(\partial_1^2+\partial_2^2)u^+(P)h^2 + O(h^3)\\ &= \partial_2 u^+(P)h \pm \partial_{12}u^+(P)h^2 + O(h^3).\end{aligned} \tag{3.6}$$

Similarly, since $\Delta u^- + \lambda u^- = 0$ in D, we obtain

$$\begin{aligned}u(Y_{\pm h}) &= -\partial_2 u^-(P)h \mp \partial_{12}u^-(P)h^2 + \frac{1}{2}(\partial_1^2+\partial_2^2)u^-(P)h^2 + O(h^3)\\ &= -\partial_2 u^-(P)h \mp \partial_{12}u^-(P)h^2 + O(h^3).\end{aligned} \tag{3.7}$$

The transmission condition (1.3) gives

$$(\partial_2 u^+(P))^2 - (\partial_2 u^-(P))^2 = \mu^2 \tag{3.8}$$

and

$$\partial_2 u^+(P)\partial_{12} u^+(P) = \partial_2 u^-(P)\partial_{12} u^-(P). \tag{3.9}$$

From (3.6), (3.7), (3.8), and (3.9), we obtain

$$|u(X_{\pm h})|^2 - |u(Y_{\pm h})|^2 = \mu^2 h^2 + O(h^4).$$

Therefore,

$$|u(X_h)|^2 - |u(Y_h)|^2 - (|u(X_{-h})|^2 - |u(Y_{-h})|^2) = O(h^4). \tag{3.10}$$

On the other hand, it follows from the maximum principle on w in (3.4), (3.5) that, for sufficiently small $h > 0$,

$$0 < |u(X_h)|^2 - |u(X_{-h})|^2 \quad \text{and} \quad 0 < |u(Y_{-h})|^2 - |u(Y_h)|^2. \tag{3.11}$$

Therefore, (3.10), (3.11) implies

$$|u(X_h)|^2 - |u(X_{-h})|^2 = O(h^4).$$

The same computation using (3.6) gives

$$|u(X_h)|^2 - |u(X_{-h})|^2 = 4h^3 \partial_2 u^+(P)\partial_{12} u^+(P) + O(h^4)$$

and we obtain $\partial_{12} u^+(P) = 0$ by comparing these two expressions. Therefore,

$$w(X_{-h}) = u(X_h) - u(X_{-h}) = O(h^3).$$

Hence, to derive a contradiction, it suffices to prove that

$$\lim_{h \to 0^+} \frac{w(X_{-h})}{h^{3-\delta}} = \infty \text{ for some } \delta > 0.$$

To do this, let us estimate w near P in a different way. We may assume $P = 0$ without loss of generality. Let us start with a truncated cone A with vertex P:

$$A := \left\{ (r, \theta) : 0 < r < r_0 \, , \; \frac{1.1}{2}\pi < \theta < \frac{1.9}{2}\pi \right\},$$

where $r_0 > 0$ is chosen so small that A is contained in $\tilde{\Sigma}^{\Omega}_{\xi}(t_0) \setminus \bar{D}$.

Define a bounded harmonic function ϕ on A as follows:

$$\phi(r, \theta) := r^{\frac{2}{0.8}} \sin\left(\frac{2}{0.8}\theta - \frac{1.1}{0.8}\pi \right).$$

Then

$$\phi(r, \frac{1.1}{2}\pi) = \phi(r, \frac{1.9}{2}\pi) = 0 \quad (0 < r < r_0).$$

Since $w > 0$ in $\tilde{\Sigma}_\xi^\Omega(t_0) \setminus \bar{D}$, $w > 0$ on $\partial A \setminus \{P\}$. Therefore, there is a positive constant C by the maximum principle so that

$$\phi(r,\theta) \leq Cw(r,\theta) \quad \text{ for all } x \in A.$$

Hence, if $0 < \delta < 3 - \frac{2}{0.8}$, then

$$\infty = \lim_{r \to 0} \frac{\phi(r,\theta)}{|r|^{3-\delta}} \leq C \lim_{r \to 0} \frac{w(r,\theta)}{|r|^{3-\delta}}.$$

This completes the proof. □

An open set Ω is said to be convex in ξ-direction if the intersection of Ω and any straight line with the direction ξ is an interval.

Corollary 3.2. *Let Ω be symmetric with respect to x_2-axis and convex in e_1-direction where e_i refers to a unit vector in the positive x_i direction. Then D is also symmetric with respect to x_2-axis and convex in e_1-direction.*

The mirror covering theorem tells us that the plasma region on each of two disks in a dumbbell shaped domain with thin and long connecting bridge is connected using mirrors passing the center points of the disks. However, the theorem does not tell us whether the number of plasma components is one or more. In fact, in some dumbbell shaped domain case presented in the paper [6], the plasma region consists of two or more components. Therefore, it is interesting to know the condition under which D is assured to be connected. We still do not know whether D is connected even in convex domain Ω; however, the following corollary provides a sufficient condition for a plasma region D to be connected.

Corollary 3.3. *Let B be a disk satisfying the requirements in Theorem* 2.2. *Suppose $\bigcup_{t>a} \tilde{\Sigma}_\xi^\Omega(t) \subset \Omega$ whenever $T_\xi(a)$ is a tangent line of ∂B. Then D is connected.*

Acknowledgment. The authors thank two anonymous referees for their careful review of the proofs in great details and also for many valuable comments to improve the manuscripts.

REFERENCES

[1] A. Acker, *On the geometric form of free boundaries satisfying a Bernoulli condition,* Math. Methods Appl. Sci., 6 (1984), pp. 449–456.

[2] H.W. Alt and L.A. Caffarelli, *Existence and regularity for a minimum problem with free boundary,* J. Reine Angew. Math., 325 (1981), pp. 105–144.

[3] H.W. Alt, L.A. Caffarelli, and Avner Friedman, *Variational problems with two phases and their free boundaries,* Trans. Amer. Math. Soc., 282 (1984), pp. 431–461.

[4] J. Athanasopoulos and L.A. Caffarelli, *A theorem of real analysis and its application to free-boundary problems,* Comm. Pure Appl. Math., 38 (1985), pp. 499–502.

[5] L.A. Caffarelli, *A Harnack inequality approach to the regularity of free boundaries,* Comm. Pure Appl. Math., 42 (1989), pp. 55–78.

[6] Avner Friedman and Yong Liu, *A free boundary problem arising in Magneto hydrodynamic system,* Ann. Scuola Norm. Sup. Pisa. Cl. Sci. (4), 22 (1995), pp. 375–448.

[7] A. Henrot and H. Shahgholian, *Convexity of free boundaries with Bernoulli type boundary condition,* Nonlinear Anal., 28 (1997), pp. 815–823.

[8] K.-K. Kang, J.-Y. Lee, and J.K. Seo, *Identification of free boundary arising in magnetohydrodynamics system,* Inverse Problems, 13 (1997), pp. 1301–1309.

[9] B. Kawohl, *Rearrangements and Convexity of Level Sets in PDE,* Lecture Notes in Math. 1150, Springer-Verlag, Berlin, New York, 1985.

[10] K.E. Lancaster, *Qualitative behavior of solutions of elliptic free boundary problems,* Pacific J. Math., 154 (1992), pp. 297–316.

[11] J. Serrin, *A symmetry problem in potential theory,* Arch. Rational Mech. Anal., 43 (1971), pp. 304–318.

[12] D.E. Tepper, *Free boundary problem,* SIAM J. Math. Anal., 5 (1974), pp. 841–846.

SIAM J. Math. Anal.
Vol. 31, No. 6, pp. 1307–1331

GLOBAL CONTINUATION VIA HIGHER-GRADIENT REGULARIZATION AND SINGULAR LIMITS IN FORCED ONE-DIMENSIONAL PHASE TRANSITIONS*

TIMOTHY J. HEALEY† AND HANSJÖRG KIELHÖFER‡

Abstract. We consider a standard "higher-gradient" model for forced phase transitions in one-dimensional, shape-memory solids. We prescribe a parameter-dependent body forcing. The component of the potential energy corresponding to conventional elasticity is characterized by a nonconvex stored energy function of the strain. Our main goal is to show that global solution branches of the regularized problem converge to a global branch of weak solutions in the limit of vanishing "capillarity" (the coefficient of the higher-gradient term). The existence of global branches for the regularized, semilinear problem is routine, based upon the Leray–Schauder degree. In the physically meaningful case when the body force is everywhere nonnegative, we obtain uniform a priori bounds via a subtle maximum principle. This together with topological connectivity arguments yields the existence of global branches of weak solutions to the zero-capillarity problem. Moreover, by examining the singular limits of various supplementary conservation laws (satisfied by all solutions of the regularized problem), we show that the above-mentioned weak solutions also minimize the potential energy of the zero-capillarity problem.

Key words. nonlinear elasticity, phase transitions, global continuation, singular limits, energy minimizer

AMS subject classifications. 34B16, 34B18, 74G25, 74G55, 74G65, 74N15

PII. S0036141098340065

1. Introduction.

Since the influential paper of Ericksen [13], the use of nonconvex stored energy functions in nonlinear elasticity as a model for martensitic phase transformations in solids is well known. In the context of one-dimensional problems, Ericksen constructed global energy minimizers possessing an arbitrary number of "phase boundaries," i.e., singular points of the associated Euler–Lagrange ordinary differential equation. In higher–dimensional models, the analogous stored energy function is allowed to violate rank-one convexity, i.e., the resulting Euler–Lagrange equations can lose ellipticity (e.g., cf. [5], [29]). Consequently, such problems are ostensibly beyond the reach of many well-known methods of nonlinear analysis—in particular, degree-theoretic methods.

Generalized degree-theoretic methods [19] were recently employed to obtain global continuation results in problems of three-dimensional nonlinear elastostatics [18]. However, the resulting solution continua are characterized not only by the usual two Rabinowitz alternatives [28], but also by the possibility that the solution branch may "terminate" due to loss of local injectivity, ellipticity, and/or the complementing condition (cf. [18, Theorem 4.1]). Physically meaningful constitutive hypotheses ruling out loss of local injectivity along global solution continua have been recently obtained in [17]. However, the potential loss of ellipticity and/or violation of the complementing condition are direct consequences of the construction of the degree. On the other

*Received by the editors June 9, 1998; accepted for publication (in revised form) September 28, 1999; published electronically May 15, 2000.

http://www.siam.org/journals/sima/31-6/34006.html

†Department of Theoretical and Applied Mechanics and Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 (tjh10@cornell.edu). The work of this author was supported by the National Science Foundation through grants DMS-9625830 and DMS-9704730.

‡Universität Augsburg, Institut für Mathematik, Universitätsstrasse 14, 86135 Augsburg, Germany.

hand, explicit examples involving homogeneous families of deformations (e.g., cf. [6], [27]) suggest that solution continua do not terminate along solution branches at the first occurrence of loss of ellipticity or failure of the complementing condition. In other words, it should be possible to obtain global branches of genuinely weak solutions in boundary value problems of nonlinear elastostatics.

In recent years many investigators have returned to original ideas of Van Der Waals [33] and Cahn and Hilliard [9] for overcoming the difficulties associated with "sharp-interface" models (loss of ellipticity). Specifically, a small term, usually quadratic in the next highest gradient of the dependent variable, is added to the energy density, which penalizes the formation of interfaces. The resulting higher-order, semilinear Euler–Lagrange equations are nonsingular, and there are many examples in the literature demonstrating the smoothening effects of small "capillarity" (the strength of the penalty term); e.g., cf. [8], [23], [26], [31], [32]. A fruitful method of analysis in such models, usually in the absence of loading, combines singular-perturbation analysis with direct methods in the calculus of variations to obtain information about energy minimizers and minimizing sequences for vanishing capillarity; cf. [8], [22], [25].

In this work we propose a different method for such problems in the presence of loading, which is complementary to the above-mentioned approach, and which potentially addresses the limitations of the global results in [18]. Namely, for parameter-dependent problems, the semilinear Euler–Lagrange equations (for small capillarity) are amenable to standard global continuation methods; cf. [28]. Our idea is to analyze the global solution continuum in the limit of vanishing capillarity, with the ultimate goal of extracting (global branches of) weak solutions of the original, zero-capillarity problem. In some sense, this is tacitly carried out in [8], [26], [23], [32] for a one-dimensional Van Der Waals problem (Ericksen's model with capillarity), where the existence of one-parameter families of solutions is obtained via phase-plane analyses. Degree-theoretic continuation methods furnish a much more general and versatile approach to existence in such problems. Finally, we mention that existence results (via the phase plane) combined with direct methods in the calculus of variations provide very detailed results in [8].

In this paper we present a model problem for which we are able to carry out the above-mentioned program in detail. The outline of the work is as follows. In section 2 we formulate a one-dimensional problem, characterized by a nonconvex stored energy function and small capillarity and subjected to a one-parameter family of body forces, the presence of the latter of which rules out the possibility of a phase-plane analysis. We then perform a standard continuation analysis via the Leray–Schauder degree (cf. [28]) to obtain global branches of solutions.

In section 3 we make the physically reasonable assumption that the body force is everywhere nonnegative. For problems characterized by single-well stored energy functions (nonconvex, with one global minimum and no other local extrema), we then show that the global solution branch of section 2 is unbounded in the Cartesian product of parameter space and an appropriate function space. Moreover, from a subtle version of the maximum principle, we deduce that every nontrivial solution on the branch is a decreasing function on its domain $[0, 1]$, which in turn enables a uniform a priori bound. In section 4 we investigate problems incorporating a double-well stored energy function (nonconvex, with precisely two local minima and one local maximum). The approach of section 3 fails in this case, since we are not able to show that the branch is unbounded, with monotonicity properties of solutions possibly being lost.

Instead, we introduce another homotopy parameter, connecting a given single-well problem to a double-well problem. We then perform a delicate continuation analysis from the single-well problem, to obtain unbounded solution branches for the double-well problem. From here the maximum principle is again employed to show that the solutions are decreasing, and we obtain a uniform a priori bound on solutions.

In section 5 we study the singular limit of solutions for vanishing capillarity. Specifically, we fix the loading parameter and consider sequences of solutions as the small penalty term goes to zero. The uniform a priori bounds and the monotonicity properties obtained in sections 3 and 4 yield detailed pointwise convergence properties (via Helly's theorem [24]) from which we can extract weak solutions of the zero-capillarity problem. Any such limiting solution is shown to possess one discontinuity in strain for large enough loading. More surprisingly, we are also able to show that the first and second Erdmann–Weierstrass corner conditions [14] (which are necessary for a strong minimum of the potential energy) are fulfilled at such a discontinuity. Together these dictate that the jump in strain occurs in accordance with the Maxwell condition. The first corner condition follows from an easy argument. The proof for the second corner condition is much more delicate; we argue via pointwise convergence of two conservation laws (identities) that are satisfied by all classical solutions of the problem with nonzero capillarity. Finally, using topological tools introduced in [1], we show that the set of all such limiting solutions (letting the loading parameter vary) comprises an unbounded branch of weak solutions—thus going beyond the limitations of [18]—albeit in this one-dimensional setting.

In section 6 we specialize to the case of dead loading and consider the stability of the limiting solutions obtained in section 5 (i.e., energy minimization properties of solutions). In this case weak solutions satisfy an algebraic equation, and for sufficiently large loading, we are able to show that the limiting continuum contains truly weak solutions. A standard application of the relaxation theorem, combined with the Maxwell condition obtained in section 5, shows that for fixed loading, each such weak solution on the limiting curve is a global energy minimizer for the zero-capillarity problem. The stability of our solutions for small, nonzero capillarity is unclear. However, in the dead-load case one can obtain global energy minimizers to the small-capillarity problem directly and show by methods similar to those of sections 4 and 5 that sequences of such minimizers also converge to the global energy minimizer of the zero-capillarity problem; cf. [3]. Whence, for sufficiently small capillarity and fixed loading, solutions from our global continua are arbitrarily "close" to the global energy minimizer.

2. Global analysis of the regularized problem.

We consider the potential energy functional

$$V_\varepsilon(\lambda, u) \equiv \int_0^1 \left[\frac{\varepsilon}{2}(u'')^2 + W(u') + B(\lambda, u, x)\right] dx \tag{2.1}$$

associated with a one-dimensional elastic medium. Here $u(x) \in \mathbb{R}$ is the displacement of the material point $x \in [0, 1]$, B is the loading potential, W is the stored energy function of classical nonlinear elasticity, $\varepsilon > 0$ is the capillarity (or higher-gradient elasticity) coefficient, and $\lambda \in \mathbb{R}$ is a loading parameter. We impose zero displacement at $x = 0$:

$$u(0) = 0. \tag{2.2}$$

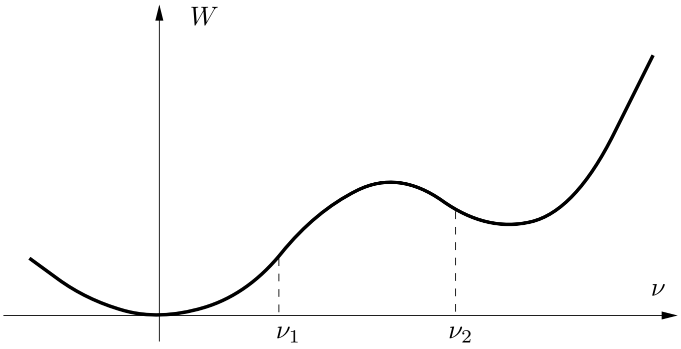


FIG. 2.1.

The Euler–Lagrange equation of equilibrium is

$$-\varepsilon u'''' + \frac{d}{dx}[W'(u')] + b(\lambda, u, x) = 0, \qquad 0 < x < 1, \tag{2.3}$$

with natural boundary conditions

$$u''(0) = u''(1) = 0, \qquad \varepsilon u'''(1) = W'(u'(1)), \tag{2.4}$$

where $b(\lambda, u, x) = -B_u(\lambda, u, x)$ is the "live" loading.

We assume that

$$\begin{aligned} &b \text{ is continuous, } b(0,\cdot,\cdot) \equiv 0,\\ &\text{and } W \text{ is of class } C^2 \text{ and nonconvex.}\end{aligned} \tag{2.5}$$

More precisely, we suppose that there are numbers $0 < \nu_1 < \nu_2$ such that

$$\begin{aligned} &W(\nu) \geq W(0) = 0 \qquad \text{for all } \nu \in \mathbb{R},\\ &\lim_{\nu\to\pm\infty} W(\nu) \to \infty,\\ &W'(0) = 0, \qquad W'(\nu) < 0, \qquad \nu \in (-\infty, 0),\\ &W''(\nu) > 0, \qquad \nu \in (-\infty, \nu_1) \cup (\nu_2, \infty),\\ &W''(\nu) < 0, \qquad \nu \in (\nu_1, \nu_2).\end{aligned} \tag{2.6}$$

A typical graph of W is depicted in Figure 2.1. We frequently denote the "stress" by $\sigma(\nu) \equiv W'(\nu)$. It is easy to see that our boundary value problem (2.2), (2.3), (2.4) is equivalent to

$$\begin{aligned} &u' = z,\\ &-\varepsilon z'' + \sigma(z) = \int_x^1 b(\lambda, u(\tau), \tau)d\tau, \qquad 0 < x < 1,\\ &u(0) = z'(0) = z'(1) = 0,\\ &\int_0^1 \sigma(z(x))dx = \int_0^1 \left[\int_x^1 b(\lambda, u(\tau), \tau)d\tau\right] dx.\end{aligned} \tag{2.7}$$

We now convert (2.7) into an operator form more convenient for global analysis.

We first define linear operators S, T as follows:

$$\begin{aligned}&w = Sf \quad \text{is the unique solution of}\\ &w' = f,\ 0 < x < 1,\ w(0) = 0\\ &\text{for } f \in C^0([0,1]).\end{aligned} \tag{2.8}$$

$$\begin{aligned}&v = Tg \quad \text{is the unique solution of}\\ &v'' = g,\ 0 < x < 1,\\ &v'(0) = v'(1) = 0, \qquad \int_0^1 v dx = 0 \quad \text{for } g \in \left\{ y \in C^0([0,1]) : \int_0^1 y dx = 0 \right\}.\end{aligned} \tag{2.9}$$

Then $S : Y \to X$ and $T : Y_1 \to Z$ are bounded, where Y, Y_1, X, Z denote the Banach spaces of continuous (continuously differentiable) functions $Y = C^0([0,1])$,

$$\begin{aligned}Y_1 &= \left\{ y \in C^0([0,1]) : \int_0^1 y dx = 0 \right\}, X = \{y \in C^1([0,1]) : y(0) = 0\},\\ Z &= \left\{ y \in C^2([0,1]) : y'(0) = y'(1) = 0, \int_0^1 y dx = 0 \right\},\end{aligned}$$

each of which is equipped with the usual supremum norm on $C^k([0,1]), k = 0, 1, 2$, denoted by $\| \cdot \|_\infty, \| \cdot \|_X, \| \cdot \|_Z$, respectively. We further define $Y_0 = \{y \in C^0[(0,1)] : y(0) = 0\}$ with norm $\| \cdot \|_\infty$.

Next we set

$$\mu = \int_0^1 z dx,\ v = z - \mu. \tag{2.10}$$

Finally we define the triple

$$\mathbf{w} \equiv (u, v, \mu) \in \mathcal{W} \equiv Y_0 \times Y_1 \times \mathbb{R}, \tag{2.11}$$

where $\mathcal{W}$ is endowed with the norm $\|\mathbf{w}\|_{\mathcal{W}} = \|u\|_\infty + \|v\|_\infty + |\mu|$. Then (2.7) is equivalent to

$$\mathbf{w} - H_\varepsilon(\lambda, \mathbf{w}) = \mathbf{0}, \tag{2.12}$$

where $H_\varepsilon : \mathbb{R} \times \mathcal{W} \to \mathcal{W}$ is defined by

$$\begin{aligned}H_\varepsilon(\lambda, \mathbf{w}) \equiv \Bigg(S(v + \mu), \frac{1}{\varepsilon}\, T \Bigg\{ &\sigma(v + \mu)\\ &- \int_x^1 b(\lambda, u, \tau) d\tau - \int_0^1 \left[\sigma(v + \mu) - \int_x^1 b(\lambda, u, \tau) d\tau \right] dx \Bigg\},\\ &\int_0^1 \left[\sigma(v + \mu) - \int_x^1 b(\lambda, u, \tau) d\tau \right] dx + \mu \Bigg).\end{aligned} \tag{2.13}$$

Since $S : Y \to Y_0$ and $T : Y_1 \to Y_1$ are compact, it readily follows that $\mathbf{w} \mapsto H_\varepsilon(\lambda, \mathbf{w})$ is compact.

From (2.5), (2.7)–(2.10), it is straightforward to verify that any solution $(\lambda, \mathbf{w}) =$

(λ, u, v, μ) delivers a classical solution $(\lambda, u, z) = (\lambda, u, v + \mu)$ of (2.7) and a classical solution (λ, u) of (2.2)–(2.4). Note that (2.7) and (2.10) give $\mu = u(1)$.

In view of (2.5), (2.6), we have

$$H_\varepsilon(0, \mathbf{0}) = \mathbf{0}, \tag{2.14}$$

and the Fréchet derivative of $\mathbf{w} \mapsto H_\varepsilon(\lambda, \mathbf{w})$ at $(0, \mathbf{0})$ is readily calculated:

$$D_{\mathbf{w}} H_\varepsilon(0, \mathbf{0})[\mathbf{h}] = \left(S(h_2 + h_3), \frac{1}{\varepsilon} W''(0) T h_2, (1 + W''(0)) h_3 \right) \tag{2.15}$$

for all $\mathbf{h} = (h_1, h_2, h_3) \in Y_0 \times Y_1 \times \mathbb{R}$. Using $W''(0) > 0$ (cf. (2.6)), it is easy to show that $[I - D_{\mathbf{w}} H_\varepsilon(0, \mathbf{0})] : \mathcal{W} \to \mathcal{W}$ is injective and thus bijective by the Riesz–Schauder theory (where "I" denotes the identity map).

Accordingly, from the implicit function theorem we have the following.

Proposition 2.1. *Equation* (2.12) *has a local curve of solutions*

$$(\lambda, \mathbf{w}) = (\lambda, \tilde{\mathbf{w}}(\lambda)), \qquad |\lambda| < \delta, \tag{2.16}$$

where $\tilde{\mathbf{w}}(0) = \mathbf{0}$. *Moreover,* (2.16) *yields all solutions of* (2.12) *in a sufficiently small neighborhood of* $(0, \mathbf{0})$.

The Leray–Schauder degree is well defined for $\mathbf{w} \mapsto \mathbf{w} - H_\varepsilon(\lambda, \mathbf{w})$, and from Proposition 2.1 we have

$$deg(I - H_\varepsilon(0, \cdot), B(\mathbf{0}), 0) = \pm 1, \tag{2.17}$$

where $B(\mathbf{0}) \subset \mathcal{W}$ denotes a sufficiently small ball centered at $\mathbf{w} = \mathbf{0}$.

A well-known argument [28], employing the homotopy invariance of the Leray–Schauder degree, yields the following.

Proposition 2.2. *Equation* (2.12) *admits a global branch of solution continua, denoted by* $\mathcal{C}_\varepsilon \subset \mathbb{R} \times \mathcal{W}$, *containing the local curve* (2.16) *and characterized by at least one of the following alternatives:*

(i) $\mathcal{C}_\varepsilon$ *is unbounded in* $\mathbb{R} \times \mathcal{W}$,

(ii) $\mathcal{C}_\varepsilon - \{(0, \mathbf{0})\}$ *is connected.*

Remark 2.3. *From* (2.5) *and* (2.7), *we note that our formulation* (2.12), (2.13) *implies* $\mathcal{C}_\varepsilon$ *is also a continuum in* $\mathbb{R} \times \tilde{Y}_o \times \tilde{Y}_1 \times \mathbb{R}$, *where* $\tilde{Y}_i = Y_i \cap C^1([0, 1])$, $i = 0, 1$, *each of which is endowed with the usual supremum norm on* $C^1([0, 1])$.

3. Detailed solution properties for single-well-potential problems.

The second alternative of Proposition (2.2) implies that problem (2.12) at $\lambda = 0$, viz., $\mathbf{w} = H_\varepsilon(0, \mathbf{w})$, admits at least one nontrivial solution $\mathbf{w} \neq \mathbf{0}$, which in turn, from (2.5), (2.9) means that the boundary value problem

$$\begin{aligned} \varepsilon z'' - \sigma(z) &= 0, \qquad 0 < x < 1, \\ z'(0) = z'(1) &= 0 \end{aligned} \tag{3.1}$$

has at least one solution $z \not\equiv 0$ in $[0, 1]$. However, if we assume that W is a *single-well potential*, i.e., in addition to (2.6), impose

$$\sigma(\nu) = W'(\nu) > 0, \qquad \nu \in (0, \infty), \tag{3.2}$$

then a standard phase-plane analysis shows that $z \equiv 0$ is the only solution of (3.1); cf. [8]. Accordingly, we have the following.

Proposition 3.1. *Under the additional hypothesis* (3.2), *the global continuum* $\mathcal{C}_\varepsilon$ *of Proposition* 2.2 *is characterized solely by alternative* (i), *i.e.*, $\mathcal{C}_\varepsilon$ *is unbounded. Moreover, if we define*

$$\mathcal{C}_\varepsilon^{+(-)} = \begin{array}{l}\textit{component of}\, \mathcal{C}_\varepsilon - \{(0,\mathbf{0})\},\\ \textit{containing}\, \{(\lambda, \tilde{\mathbf{w}}(\lambda)) : 0 < \lambda < \delta(-\delta < \lambda < 0)\},\end{array} \tag{3.3}$$

we then have the disjoint union $\mathcal{C}_\varepsilon = \mathcal{C}_\varepsilon^+ \cup \{(0,\mathbf{0})\} \cup \mathcal{C}_\varepsilon^-$, *where* $\mathcal{C}_\varepsilon^+$ *and* $\mathcal{C}_\varepsilon^-$ *are each unbounded in* $\mathbb{R} \times \mathcal{W}$.

Remark 3.2. *We suspend hypothesis* (3.2) *later in section* 4.

Next we define

$$P \equiv \{h \in C^1([0,1]) : h < 0 \text{ in } (0,1), h(0) = h(1) = 0,\ h'(0) < 0,\ h'(1) > 0\}. \tag{3.4}$$

Throughout the remainder of this work, we further assume that the body force satisfies the following physically realistic assumption (e.g., gravitation):

$$\begin{array}{ll} b(\lambda, u(x), x) \geq 0, \qquad 0 \leq x \leq 1, \qquad \text{and} \\ b(\lambda, \cdot) \not\equiv 0 \quad \text{for all } \lambda > 0, u \in Y_0. \end{array} \tag{3.5}$$

We then obtain the following.

Theorem 3.3. *Under the hypotheses of section* 1 *and the additional assumptions* (3.2) *and* (3.5), *it follows that* $(\lambda, u, v, \mu) \in \mathcal{C}_\varepsilon^+$ *implies*

$$z' = v' \in P.$$

Proof. First we show that $\frac{d}{dx}\tilde{z}(\lambda) = \frac{d}{dx}(\tilde{v}(\lambda) + \tilde{\mu}(\lambda)) \in P$ for all $0 < \lambda < \delta$, for δ sufficiently small; cf. (2.16) and (3.3). Indeed, for $0 < \lambda < \delta$, we have $\|\tilde{z}(\lambda)\|_\infty \leq \|\tilde{v}(\lambda)\|_\infty + |\tilde{\mu}(\lambda)| < \gamma(\delta)$, with $\gamma(\delta) \searrow 0$ as $\delta \searrow 0$. Hence, from $(2.6)_4$ we find

$$\sigma'(\tilde{z}(\lambda)) = W''(\tilde{z}(\lambda)) > 0 \quad \text{for all } x \in [0,1], \tag{3.6}$$

with δ sufficiently small. Differentiation of $(2.7)_2$ and $(2.7)_3$ shows that $h = \frac{d}{dx}\tilde{z}(\lambda)$ is a solution of

$$\begin{aligned} \varepsilon h'' - [\sigma'(\tilde{z}(\lambda))]h &= b(\lambda, \tilde{u}(\lambda), \cdot), \qquad 0 < x < 1, \\ h(0) = h(1) &= 0. \end{aligned}$$

In view of (3.5), (3.6), and the maximum principle, we conclude that $\frac{d}{dx}\tilde{z}(\lambda) \in P$ for all $0 < \lambda < \delta$.

Since P is open in $C^1([0,1])$, and $\frac{d}{dx}\tilde{z}(\lambda) \in P$, $0 < \lambda < \delta$, it follows that if $z' = v' \notin P$ for $(\lambda, u, v, \mu) \in \mathcal{C}_\varepsilon^+$, then there is a sequence $\{(\lambda_j, u_j, v_j, \mu_j)\} \subset \mathcal{C}_\varepsilon^+$ such that $(\lambda_j, u_j, v_j, \mu_j) \to (\lambda, u, v, \mu)$ in $\mathbb{R} \times \tilde{Y}_0 \times \tilde{Y}_1 \times \mathbb{R}$ (cf. Remark 2.3), with $z_j' = v_j' \in P$ and $z' = v' \in \partial P$, i.e.,

$$z' \leq 0 \text{ in } (0,1) \text{ and/or } z''(0) = 0 \text{ and/or } \ z''(1) = 0. \tag{3.7}$$

But again by (2.7), we see that $h = z'$ is a solution of

$$\begin{array}{c} \varepsilon h'' - [\sigma'(z)]^+ h = b(\lambda, u, \cdot) + [\sigma'(z)]^- z', \qquad 0 < x < 1, \\ h(0) = h(1) = 0, \end{array} \tag{3.8}$$

where for any continuous function $p(x)$ on $[0,1]$,

$$p(x)^{+(-)} \equiv \begin{cases} p(x) & \text{for all } x \in [0,1] \text{ such that } p(x) \geq (\leq)\, 0, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, the right-hand side of (3.8) is nonnegative and does not vanish identically; cf. (3.5), (3.7). Accordingly, (3.7) contradicts the maximum principle ($z' \equiv 0$ is not possible, due to (3.5)), and thus $z' \in P$. □

With Theorem 3.3 in hand, we can now obtain an a priori bound on $\mathcal{C}_\varepsilon^+$ that is crucial to our forthcoming limit analysis in section 5, provided that we make a final assumption on the loading:

$$b(\lambda, u(x), x) \leq \tilde{b}(\lambda, x) \tag{3.9}$$

for all $\lambda > 0, u \in Y_0$, where $\tilde{b}$ is some continuous, nonnegative function; cf. (3.5).

Remark 3.4. *Note that* (3.9) *is automatic in the case of "dead loading," i.e., when b is independent of u. We return to this special case in section* 6.

Theorem 3.5. *Given the hypotheses of Theorem* 3.3 *and assumption* (3.9)*, then for any* $(\lambda, u, v, \mu) \in \mathcal{C}_\varepsilon^+$, *with* $z = v+\mu$, *there is a positive constant* $M(\lambda)$, *independent of* ε, *such that*

$$\|z\|_\infty \leq M(\lambda), \tag{3.10}$$

where $M(\lambda)$ *depends upon* λ *through* $F(\lambda) = \int_0^1 \tilde{b}(\lambda, x)dx$ *and the graph of* $\sigma = W'$. *Moreover,*

$$\|u\|_{C^1([0,1])} \leq 2M(\lambda). \tag{3.11}$$

Accordingly, the projection of $\mathcal{C}_\varepsilon^+$ *on the* λ*-axis is* $(0, \infty)$, *i.e.,* (2.12) *(and thus,* (2.2)–(2.4) *or* (2.7)*) has at least one solution for each* $\lambda \in (0, \infty)$.

Proof. From (2.4), (2.7), and Theorem 3.3, we see that

$$\begin{aligned} &z' < 0, \qquad 0 < x < 1, \\ &\sigma(z(1)) = \varepsilon z''(1) > 0, \\ &-\varepsilon z''(0) > 0. \end{aligned} \tag{3.12}$$

In view of (2.6) and (3.2), a typical graph of σ is depicted in Figure 3.1. In particular, $(3.12)_2$ implies that $z(1) > 0$, and thus $(3.12)_1$ yields

$$\|z\|_\infty = z(0). \tag{3.13}$$

Finally, $(2.7)_2$ combined with (3.9) and $(3.12)_3$ gives

$$\sigma(z(0)) < \int_0^1 b(\lambda, u(x), x)dx \leq \int_0^1 \tilde{b}(\lambda, x)dx = F(\lambda), \tag{3.14}$$

where $F(\lambda) > 0$. Then (3.10) follows from (3.13) and (3.14) with $M(\lambda) \equiv \max\{\nu : \sigma(\nu) = F(\lambda)\}$; cf, Figure 3.1. Inequality (3.11) is an easy consequence of $(2.7)_1$, $(2.7)_3$, and (3.10). From the development above, we have $z > 0$ and monotone decreasing on $[0,1]$. Then by virtue of (2.10) and (3.13), we see that the positive numbers $v(0)$, $|v(1)|$ and μ are each bounded above by $\|z\|_\infty$. Thus, by (2.7) and (3.10), we conclude

$$\|\mathbf{w}\|_{\mathcal{W}} = \|u\|_\infty + \|v\|_\infty + |\mu| \leq 3M(\lambda), \tag{3.15}$$

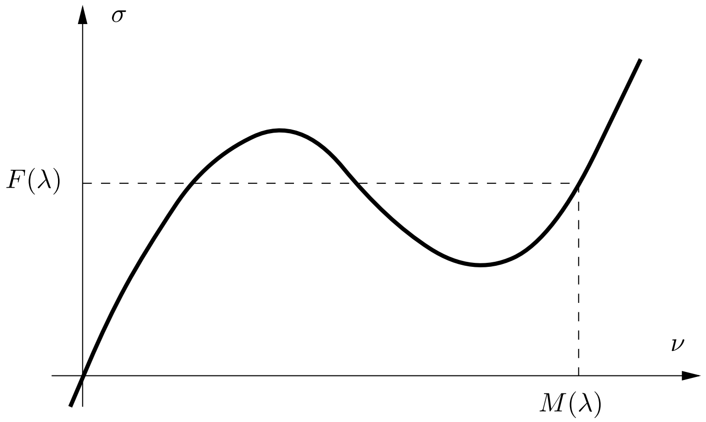


Fig. 3.1.

for all $(\lambda, \mathbf{w}) \in \mathcal{C}_\varepsilon^+$. Since $\mathcal{C}_\varepsilon^+$ is unbounded, the last claim follows directly from (3.15). □

Remark 3.6. *Note that* $b(0,\cdot,\cdot) \equiv 0$ *implies* $M(0) = 0$, *which is consistent with the fact that* $z \equiv 0$ *is the only solution of* (3.1).

4. Detailed solution properties for double-well-potential problems. In the absence of assumption (3.2), in which case we call $W(\nu)$ a *double-well potential* (cf. Figure 2.1), we cannot directly obtain Theorems 3.3 and 3.5 as in section 3. Indeed, without (3.2), we do not have Proposition 3.1, in which case the solution continuum $\mathcal{C}_\varepsilon$ could form a bounded "loop," cf. Proposition 2.2 (ii). Moreover, $\mathcal{C}_\varepsilon^+$ (cf. (3.3)) may have components in the "half-space" $(-\infty, 0) \times \mathcal{W}$. Thus the positivity of b could be lost (cf. (3.5)), in which case the maximum principle arguments in the proof of Theorem 3.3 fail. Of course the results of Theorem 3.3 and Theorem 3.5 still hold along part of $\mathcal{C}_\varepsilon^+$, which we record now for convenience. Define $\mathcal{L}_\varepsilon^+ \subset \mathcal{C}_\varepsilon^+ \cap ((0,\infty) \times \mathcal{W})$ to be the connected subset such that $\overline{\mathcal{L}_\varepsilon^+}$ contains the point $(\lambda, \mathbf{w}) = (0, \mathbf{0})$. We then have the following.

Corollary 4.1. *Assume that* $W(\nu)$ *is a double-well potential (i.e., suspend assumption* (3.2)*). Then the results of Theorem* 3.3 *and Theorem* 3.5 *hold for all* $(\lambda, \mathbf{w}) \in \mathcal{L}_\varepsilon^+$.

Remark 4.2. *Returning to the proof of Theorem* 3.5, *note that in this case the graph of* $\sigma = W'$ *intersects the positive* ν*-axis; cf. Figure* 4.1. *Whence,* $M(0) > 0$ *for double-well potentials; in particular,* (3.1) *has more than one constant solution. Consequently,* $\mathcal{L}_\varepsilon^+$ *could be bounded. Thus, in contrast to the results of section* 3, *we do not yet know if problem* (2.7) *has a solution for a given* $\lambda > 0$ *when* $W(\nu)$ *is a two-well potential; cf.* Theorem 3.5.

In the remainder of this section we establish the existence of an unbounded solution set for (2.7) $(\lambda > 0)$, with properties similar to those given in Theorems 3.3 and 3.5, in the absence of assumption (3.2). To begin, let $W_1(\nu)$ be a potential satisfying (2.6) and (3.2), while $W_2(\nu)$ is another potential that satisfies conditions (2.6) but not (3.2); i.e., $W_1(\nu)$ has a "single well," while $W_2(\nu)$ has a "double well." We then define the homotopy

$$\tilde{W}(\nu, \tau) \equiv (1-\tau)W_1(\nu) + \tau W_2(\nu), \qquad 0 \le \tau \le 1. \tag{4.1}$$

The basic idea is to substitute $\tilde{W}(\nu, \tau)$ into our formulation from section 2. At $\tau = 0$,

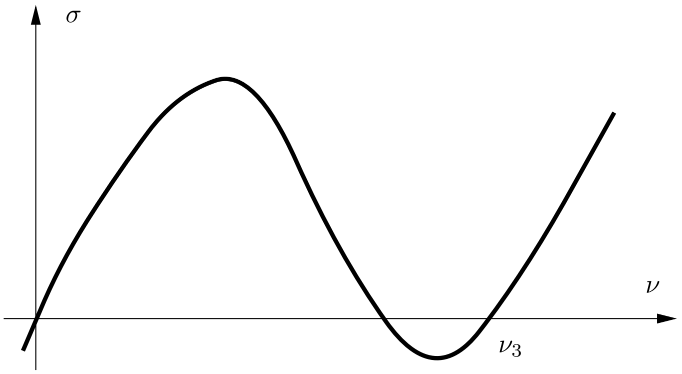


FIG. 4.1.

we then have the results of section 3; we now show that continuation in "τ" yields solutions with similar properties at $\tau = 1$.

Substitution of $\tilde{\sigma}(\nu, \tau) = \tilde{W}_\nu(\nu, \tau)$ into (2.12), (2.13) yields a two-parameter problem, which we henceforth denote by

$$\mathbf{w} - \tilde{H}_\varepsilon(\tau, \lambda, \mathbf{w}) = 0, \tag{4.2}$$

where $\tilde{H}_\varepsilon : \mathbb{R} \times \mathbb{R} \times \mathcal{W} \to \mathcal{W}$; cf. (2.11). Clearly

$$\tilde{H}_\varepsilon(0, \lambda, \mathbf{w}) \equiv H_\varepsilon(\lambda, \mathbf{w}), \tag{4.3}$$

(cf. (2.13), (4.1)), while (4.3) at $\tau = 1$ represents a given two-well-potential problem, solutions of which we hope to construct.

First we define a continuous mapping $\hat{H}_\varepsilon : \mathbb{R} \times \mathcal{W} \to \mathcal{W}$, as follows:

$$\hat{H}_\varepsilon(\theta, \mathbf{w}) \equiv \begin{cases} \tilde{H}_\varepsilon(0, \theta, \mathbf{w}), & \theta \le \lambda_0, \\ \tilde{H}_\varepsilon(\theta - \lambda_0, \lambda_0, \mathbf{w}), & \lambda_0 \le \theta \le \lambda_0 + 1, \\ \tilde{H}_\varepsilon(1, \theta - 1, \mathbf{w}), & \theta \ge \lambda_0 + 1, \end{cases} \tag{4.4}$$

where $\lambda_0 > 0$ is fixed. We then consider the problem

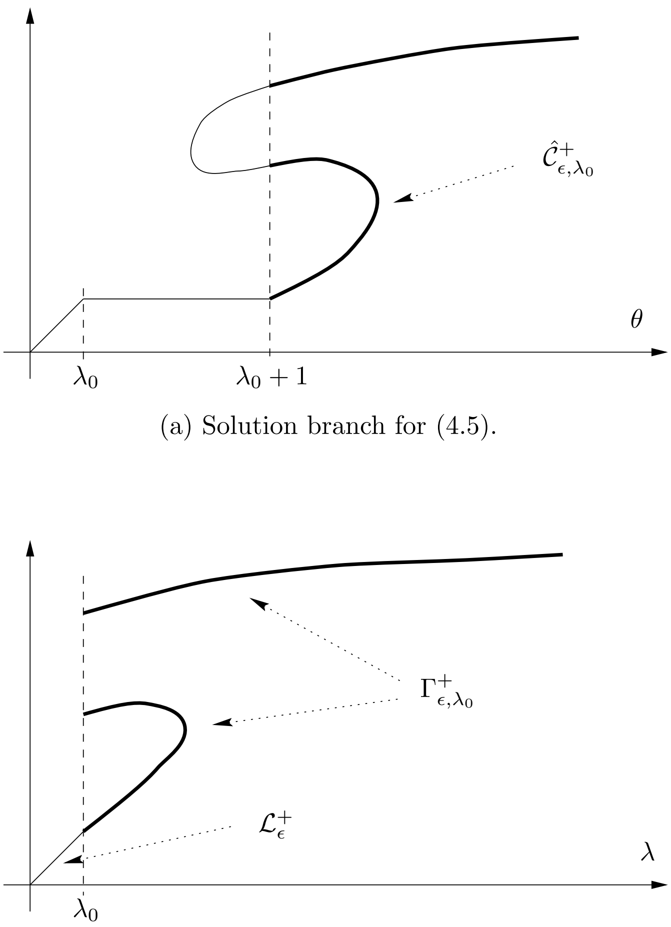
$$\mathbf{w} - \hat{H}_\varepsilon(\theta, \mathbf{w}) = \mathbf{0}. \tag{4.5}$$

It is easy to check that the mapping $\hat{H}_\varepsilon$ and problem (4.5) fulfill the hypotheses of Propositions 2.2 and 3.1. Accordingly, we have the following.

PROPOSITION 4.3. *Equation* (4.5) *admits an unbounded branch of solutions containing the point* $(0, \mathbf{0})$, *denoted* $\hat{\mathcal{C}}_{\varepsilon,\lambda_0} \subset \mathbb{R} \times \mathcal{W}$. *In addition, if we define* $\hat{\mathcal{C}}^{+(-)}_{\varepsilon,\lambda_0}$ *as in* (3.3), *then we have the disjoint union* $\hat{\mathcal{C}}_{\varepsilon,\lambda_0} = \hat{\mathcal{C}}^{+}_{\varepsilon,\lambda_0} \cup \{(0, \mathbf{0})\} \cup \hat{\mathcal{C}}^{-}_{\varepsilon,\lambda_0}$, *where* $\hat{\mathcal{C}}^{+}_{\varepsilon,\lambda_0}$ *and* $\hat{\mathcal{C}}^{-}_{\varepsilon,\lambda_0}$ *are each unbounded.*

Next we continue to retrace our steps from section 3.

THEOREM 4.4. *Given assumptions* (3.2) *(for* $\tilde{W}_\nu(\nu, 0) = \tilde{\sigma}(\nu, 0)$*) and* (3.5), *it follows that* $(\theta, u, v, \mu) \in \hat{\mathcal{C}}^{+}_{\varepsilon,\lambda_0}$, *with* $z = v + \mu$, *implies* $z' = v' \in P$; *cf.* (3.4). *Moreover, we again have uniform (independent of* ε*) bounds*



(a) Solution branch for (4.5).

(b) Solution branch for (4.2) at $\tau = 1$.

FIG. 4.2.

$$\begin{aligned} &\|z\|_\infty \leq M(\lambda) \\ (4.6) \qquad &\textit{and} \\ &\|u\|_{C^1([0,1])} \leq 2M(\lambda), \end{aligned}$$

where $M(\lambda)$ *is defined as in the proof of Theorem* 3.5 *for* $0 < \lambda = \theta \leq \lambda_0$ *(with* $\tau = 0$*) and also for* $\lambda = \theta - 1 \geq \lambda_0$ *(with* $\tau = 1$*). For* $\lambda = \lambda_0$*, and* $\lambda_0 \leq \theta \leq \lambda_0 + 1$*,* M *depends upon* θ *via*

$$M(\theta) = \max_\nu \{\nu : \tilde{\sigma}(\nu, \theta - \lambda_0) = F(\lambda_0)\}, \qquad \textit{where } F(\lambda) \textit{ is as defined in } (3.14).$$

Proof. The proof is identical to the combined proofs of Theorem 3.3 and 3.5, except that the analogue of (3.14) now reads

$$\begin{array}{ll} \sigma(z(0), 0) < F(\lambda), & 0 < \lambda = \theta \leq \lambda_0, \\ \sigma(z(0), \theta - \lambda_0) < F(\lambda_0), & \lambda_0 \leq \theta \leq \lambda_0 + 1, \\ \sigma(z(0), 1) < F(\lambda), & \lambda = \theta - 1 \geq \lambda_0. \end{array}$$

Estimates (4.6) then follow as before from the graph of $\tilde{\sigma}$; cf. Figure 3.1. $\square$

By construction, the solution sets

$$\Gamma^-_{\varepsilon,\lambda_0} \equiv \hat{\mathcal{C}}_{\varepsilon,\lambda_0} \cap ((-\infty,\lambda_0] \times \mathcal{W}) \equiv \mathcal{C}_\varepsilon \cap ((-\infty,\lambda_0] \times \mathcal{W})$$

and

$$\Gamma^+_{\varepsilon,\lambda_0} \equiv \hat{\mathcal{C}}^+_{\varepsilon,\lambda_0} \cap ([\lambda_0+1,\infty) \times \mathcal{W}), \tag{4.7}$$

which solve (4.2) at $\tau = 0$ and at $\tau = 1$, respectively, are unbounded. In particular, the estimates (4.6) insure that the double-well problem ($\tau = 1$) has at least one solution for every value of $\lambda \geq \lambda_0$.

REMARK 4.5. $\Gamma^+_{\varepsilon,\lambda_0}$ *need not be a continuum; it could comprise a union of disjoint continua. To see this, suppose that* $\mathcal{L}^+_\varepsilon$ *is bounded, and choose* $\lambda_0 > 0$ *sufficiently small. An easy argument via the implicit function theorem shows that* $\mathcal{L}^+_\varepsilon \cap ([\lambda_0,\infty) \times \mathcal{W}) \subset \Gamma^+_{\varepsilon,\lambda_0}$. *Since* $\Gamma^+_{\varepsilon,\lambda_0}$ *is unbounded, it must have an unbounded disjoint component. We illustrate this schematically in Figure* 4.2.

5. Singular limit analysis. In this section we demonstrate that the classical solutions of (2.2)–(2.4) ($\varepsilon > 0$) obtained in the previous sections converge to a global continuum of weak solutions of (2.3) ($\varepsilon = 0$) in the limit as $\varepsilon \searrow 0$. For the one-well problem of section 3, we fix $\lambda \in (0,\infty)$ with $\varepsilon = \varepsilon_k$, $\varepsilon_k \searrow 0$, we denote a classical solution of (2.2)–(2.4) from $\mathcal{C}^+_{\varepsilon_k}$ by u_k, and we consider sequences $\{u_k\} \subset C^1([0,1])$ and $\{z_k\} = \{u_k'\} \subset C^0([0,1])$. For double-well problems (cf. section 4), we fix $\theta \in (0,\infty)$ and consider sequences of solutions from $\hat{\mathcal{C}}^+_{\varepsilon_k,\lambda_0}$ as above.

By virtue of either (3.10), (3.11), or (4.6), we have the uniform bounds

$$\begin{aligned} \|z_k\|_\infty &\leq M(\lambda) \\ &\text{and} \\ \|u_k\|_{C^1([0,1])} &\leq 2M(\lambda) \quad \text{for all } k \in \mathbb{N}. \end{aligned} \tag{5.1}$$

From the latter and compact embedding, we conclude immediately that

$$u_k \to u_* \quad \text{in} \qquad C^0([0,1]). \tag{5.2}$$

In addition to $(5.1)_1$, we have from either Theorem 3.3 or Theorem 4.4 the monotonicity property

$$z_k' < 0, \qquad 0 < x < 1 \quad \text{for all } k \in \mathbb{N}. \tag{5.3}$$

By Helly's (selection principle) theorem (cf. [24, section 24]), it then follows that $\{z_k\}$ has a subsequence that converges pointwise on $[0,1]$ to a nonincreasing function, denoted

$$z_k \to z_*, \tag{5.4}$$

where z_* has at most a countable number of discontinuities on $[0,1]$. Moreover, by Lebesgue's differentiation theorem [30], it follows that z_* possesses a classical derivative almost everywhere (a.e.); in particular

$$z_*' \leq 0 \quad \text{a.e. in} \quad [0,1]. \tag{5.5}$$

We also note from (2.7), (5.1), and Lebesgue's dominated convergence theorem that

$$\begin{aligned} u_*(x) &= \int_0^x z_*(\tau)d\tau, \qquad \text{and hence} \\ u_*' &= \ z_* \quad \text{a.e.} \end{aligned} \tag{5.6}$$

Next we show that u_* satisfies (2.3) with $\varepsilon = 0$.

THEOREM 5.1. *The limit u_* is a weak solution of* (2.3) *for* $\varepsilon = 0$ (*with* $\lambda \in (0, \infty)$ *fixed*). *Moreover, $u_*' = z_*$ has at most one "jump" discontinuity at, say, $x = a$ and possesses a classical derivative for all $x \neq a$. In particular,* (5.5) *holds pointwise on* $[0, a) \cup (a, 1]$.

Proof. We evaluate (2.7) at ε_k, u_k, z_k, multiply by a test function φ, and integrate by parts to obtain

$$-\int_0^1 \varepsilon_k z_k \varphi'' = \int_0^1 \left[-\sigma(z_k) + \int_x^1 b(\lambda, u_k(\tau), \tau) d\tau \right] \varphi dx \tag{5.7}$$

for all $\varphi \in C_0^\infty(0, 1)$. In view of (5.1), the continuity of σ and b, and the dominated convergence theorem, the limit of (5.7) as $k \to \infty$ yields

$$0 = \int_0^1 \left[-\sigma(z_*) + \int_x^1 b(\lambda, u_*(\tau), \tau) d\tau \right] \varphi dx \tag{5.8}$$

for all test functions φ, which proves

$$\sigma(z_*(x)) = \int_x^1 b(\lambda, u_*(\tau), \tau) d\tau \quad \text{for all} \quad x \in [0, 1]. \tag{5.9}$$

Then $(5.6)_1$, (5.9) and the continuity of b yield

$$\sigma \circ z_* \in C^1[0, 1]. \tag{5.10}$$

Moreover, (3.5) and (5.9) show that $\sigma \circ z_*$ is monotone decreasing on $[0, 1]$. Hence, (5.5) and the graph of σ (cf. Figures 3.1, 4.1) then show that z_* can suffer at most one jump discontinuity. Finally, if we "invert" σ in (5.9) (on the two branches of z_*, for which σ is monotone increasing), we conclude that z_* is C^1 on $[0, a) \cup (a, 1]$. □

COROLLARY 5.2. *The limit u_* satisfies the Euler–Lagrange equation* (*for* (2.1) *with* $\varepsilon = 0$);

$$W''(u_*')u_*'' + b(\lambda, u_*, x) = 0 \quad \text{a.e. in } [0, 1]. \tag{5.11}$$

Proof. Differentiation of (5.9) (cf. (5.10)), using (5.6) and the differentiability of z_* (cf. (5.5)), yields (5.11). □

Our calculations thus far leave the location of a possible jump, say, at $x = a$, indeterminate. This is easy to see, e.g., for dead loading, in which case (5.9) reduces to

$$\sigma(z_*(x)) = \int_x^1 b(\lambda, \tau) d\tau \equiv F(\lambda, x). \tag{5.12}$$

Referring to Figure 3.1, we see that z_* can have at most one jump (again by monotonicity) but that there are infinitely many choices for $\sigma(z_*(a^+)) = \sigma(z_*(a^-))$.

(In this case, any value between the local minimum value and the local maximum value of σ will do.) We now provide an additional argument to show that the jump in z_* must occur in accordance with the second Erdmann–Weierstrass corner condition.

Theorem 5.3. *The limiting strain* $u'_* = z_*$ *satisfies the second corner condition:*

$$W(z_*) - z_*\sigma(z_*) \quad \text{is continuous on } [0,1]. \tag{5.13}$$

Proof. If z_* is continuous on $[0,1]$, then (5.13) is obvious. Accordingly, we assume that z_* has a single "jump" discontinuity at $x = a$, where $0 < a \le 1$. (Note that $a = 1$ is possible only for double-well potentials, since in that case σ is no longer positive on all of $(0, \infty)$; cf. Figure 4.1.) Consider the sequences $\{u_k\}, \{z_k = u'_k\}$ for fixed $\lambda \in (0,\infty)$ or fixed $\theta \in (0,\infty)$ as before. Observe that any classical solution u_k of (2.3) with $\varepsilon = \varepsilon_k$ satisfies the Noether-type conservation law

$$\begin{aligned} W(z_k) - z_k\sigma(z_k) = \tfrac{1}{2}\varepsilon_k(z'_k)^2 - \varepsilon_k z''_k z_k \\ + \int_0^x b(\lambda, u_k(\tau), \tau) z_k(\tau) d\tau + c_k, \end{aligned} \tag{5.14}$$

where $c_k = W(z_k(0)) - z_k(0)\sigma(z_k(0)) + \varepsilon_k z''_k(0) z_k(0)$. (One can easily verify (5.14) via direct differentiation. In the absence of the inhomogeneous body force, one obtains (5.14) from the invariance of (2.3) under $x \to x - \alpha$ via Noether's theorem.) The idea is to take the pointwise limit of (5.14) and show that the left-hand side is continuous as $k \to \infty$. We now split a large part of the proof into the following two lemmas.

Lemma 5.4. $\lim_{k\to\infty} \varepsilon_k z''_k z_k = 0$ *for every* $x \in [0,1]$.

Proof. From $(2.7)_2$ we have

$$\varepsilon_k z''_k z_k = \left[\sigma(z_k) - \int_x^1 b(\lambda, u_k(\tau), \tau) d\tau\right] z_k. \tag{5.15}$$

Using (5.1), (5.2), (5.4), and the dominated convergence theorem, we find that the right-hand side of (5.15) converges pointwise to $[\sigma(z_*) - \int_x^1 b(\lambda, u_*(\tau), \tau) d\tau] z_*$ on $[0,1]$. Hence (5.9) gives the desired result. □

Lemma 5.5.

$$\lim_{k\to\infty} \varepsilon_k (z'_k)^2 = 0 \ \textit{a.e. on } [0,1]. \tag{5.16}$$

Proof. We first establish (5.16) on $[0,a)$. From $(2.7)_2$ we note that any classical solution pair (u_k, z_k) satisfies the Hamiltonian-type conservation law

$$\begin{aligned} \tfrac{1}{2}\varepsilon_k(z'_k)^2 = & \left[W(z_k(\cdot)) - \left(\int_{(\cdot)}^1 b(\lambda, u_k(\tau), \tau) d\tau\right) z_k\right]\Bigg|_0^x \\ & - \int_0^x b(\lambda, u_k(\tau), \tau) z_k(\tau) d\tau. \end{aligned} \tag{5.17}$$

(Equation (5.17) is readily verified from $(2.7)_2$ via direct integration. If one views "x" as a time-like variable in $(2.7)_2$, then (5.17) expresses balance of "energy".) By the dominated convergence theorem, the right-hand side of (5.17) converges pointwise on $[0,a)$ to

$$\begin{aligned} & \left[W(z_*(\cdot)) - \left(\int_{(\cdot)}^1 b(\lambda, u_*(\tau), \tau) d\tau\right) z_*\right]\Bigg|_0^x \\ & - \int_0^x b(\lambda, u_*(\tau), \tau) z_*(\tau) d\tau. \end{aligned} \tag{5.18}$$

On the other hand, (5.9) and the properties of z_* imply that

$$\sigma(z_*)z_*' = \left(\int_x^1 b(\lambda, u_*(\tau), \tau)d\tau\right) z_*' \quad \text{pointwise on } [0,a) \cup (a,1]. \tag{5.19}$$

Integration of (5.19) from 0 to x ($x < a$) and integration by parts shows that (5.18) vanishes identically on $[0,a)$; i.e., (5.16) holds pointwise on $[0,a)$. If $a = 1$, then we are finished. If $a < 1$, then we consider a different version of (5.17), viz.,

$$\begin{aligned} \frac{1}{2}\varepsilon_k(z_k')^2 = &- \left[W(z_k(\cdot)) - \left(\int_{(\cdot)}^1 b(\lambda, u_k(\tau), \tau)d\tau\right) z_k\right]\Bigg|_x^1 \\ &+ \int_x^1 b(\lambda, u_k(\tau), \tau)z_k(\tau)d\tau. \end{aligned} \tag{5.20}$$

In the limit as $k \to \infty$ of (5.20), the right-hand side converges pointwise on $(a,1]$ to the same terms with "$*$" in place of "k." On the other hand, integration of (5.19) from x to 1 ($x > a$) and integration by parts shows that the right-hand side of (5.20) with "$k = *$" vanishes identically on $(a,1]$. □

Proof of Theorem 5.3 (continued). By (5.1) the proof of Lemma 5.4 shows also that the sequence $(\varepsilon_k z_k'' z_k)$ is pointwise bounded on $[0,1]$. We now take the pointwise limit $k \to \infty$ of (5.14) (passing, if necessary, to a subsequence such that $\varepsilon_k z_k''(0)z_k(0)$ converges), using Lemmas 5.4 and 5.5, which yields

$$W(z_*) - z_*\sigma(z_*) = \int_0^x b(\lambda, u_*(\tau), \tau)d\tau + c_*, \tag{5.21}$$

where $c_* = W(z_*(0)) - z_*(0)\sigma(z_*(0))$. □

The continuity of $\sigma(z_*)$ (cf. (5.10)), often referred to as the first Erdmann–Weierstrass corner condition, combined with (5.13) yields the Maxwell "equal-area" rule for $\sigma(z_*(a^+)) = \sigma(z_*(a^-))$:

$$W(z_*(a^-)) - W(z_*(a^+)) = \sigma(z_*(a^-))\left[z_*(a^-) - z_*(a^+)\right]. \tag{5.22}$$

The two corner conditions are both necessary for $V_0(\lambda, u)$ (cf. (2.1)) to have a minimum at u_*. We return to this issue in section 6 in the context of dead loading.

Our last result in this section is to show that the limiting set of weak solutions of (2.3) ($\varepsilon = 0$) for $\lambda \geq 0$ is a continuum. Here we directly employ the point-set topological arguments of [1]. We first observe that (5.1) and (5.4) imply $\|z_*\|_\infty \leq M(\lambda)$, and thus, in view of (5.6), we conclude (for fixed $\lambda > 0$ or fixed $\theta > 0$)

$$\begin{aligned} &z_k \to z_* = u_*' \quad \text{in } L^p(0,1) \text{ and} \\ &u_k \to u_* \quad \text{in } W^{1,p}(0,1) (p \geq 1), \end{aligned} \tag{5.23}$$

where $\varepsilon = \varepsilon_k \searrow 0$ as before. Accordingly, for single-well problems we define

$$\begin{aligned} \Sigma_0^+ = \Big\{&(\lambda, u) \in \mathbb{R} \times W^{1,p}(0,1) : (\lambda_k, u_k, v_k, \mu_k) \in \overline{\mathcal{C}_{\varepsilon_k}^+}, \\ &\lambda_k \to \lambda, u_k \to u \text{ in } W^{1,p}(0,1) \text{ as } \varepsilon_k \searrow 0\Big\}. \end{aligned} \tag{5.24}$$

For double-well problems, we define

$$\begin{aligned}\Sigma^+_{0,\lambda_0} = \Big\{(\theta, u) \in \mathbb{R} \times W^{1,p}(0,1) : (\theta_k, \nu_k, \mu_k) \in \overline{\hat{\mathcal{C}}^+_{\varepsilon_k\lambda_0}},\\ \theta_k \to \theta, u_k \to u \text{ in } W^{1,p}(0,1) \text{ as } \varepsilon_k \to \infty\Big\}.\end{aligned} \tag{5.25}$$

For $\varepsilon > 0$, it is convenient to introduce the solution sets

$$\begin{aligned}&\Sigma^+_{\varepsilon} = \Big\{(\lambda, u) : (\lambda, u, v, \mu) \in \overline{\mathcal{C}^+_{\varepsilon}}\Big\}, && \text{(a)}\\ &\Sigma^+_{\varepsilon,\lambda_0} = \Big\{(\theta, u) : (\theta, u, v, \mu) \in \overline{\mathcal{C}^+_{\varepsilon,\lambda_0}}\Big\}, && \text{(b)}\end{aligned} \tag{5.26}$$

which like $\overline{\mathcal{C}^+_{\varepsilon}}$ and $\overline{\hat{\mathcal{C}}^+_{\varepsilon,\lambda_0}}$, respectively, comprise continua; i.e., Σ^+_{ε} and $\Sigma^+_{\varepsilon,\lambda_0}$ are each connected and (locally) compact.

Next let $B_R \subset W^{1,p}(0,1)$ denote the closed ball of radius "R," centered at the origin. For any number $\gamma_0 > 0$, we consider the sets

$$A = \{0\} \times B_{2M(0)} \quad \text{and} \quad B = \{\gamma_0\} \times B_{2M(\gamma_0)}, \tag{5.27}$$

where "M" is the constant (independent of ε) in the estimates (5.1). Both A and B are closed subsets of the Banach space $\mathbb{R} \times W^{1,p}(0,1)$. Henceforth we define the subsets $\Upsilon_{\varepsilon,\gamma_0}$ via either

$$\Upsilon^+_{\varepsilon,\gamma_0} \equiv \Sigma^+_{\varepsilon} \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big)$$

or

$$\Upsilon^+_{\varepsilon,\gamma_0} \equiv \Sigma^+_{\varepsilon,\lambda_0} \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big), \tag{5.28}$$

depending on whether we are interested in solutions from $\mathcal{C}^+_{\varepsilon}$ or $\hat{\mathcal{C}}^+_{\varepsilon,\lambda_0}$, respectively. The following construction is valid in either case. By (5.1) we have

$$\begin{aligned}&A \cap \Upsilon^+_{\varepsilon_k,\gamma_0} \neq \emptyset \text{ and } B \cap \Upsilon^+_{\varepsilon_k,\gamma_0} \neq \emptyset\\ &\text{are not separated in } \Upsilon^+_{\varepsilon_k,\gamma_0} \text{ for any } \varepsilon_k \searrow 0.\end{aligned} \tag{5.29}$$

To complete the argument, we need the following.

LEMMA 5.6. *For each neighborhood N of $\Sigma^+_0 \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big)$ (or of $\Sigma^+_{0,\lambda_0} \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big)$, there is a $k \in \mathbb{N}$ such that $\Upsilon^+_{\varepsilon_k,\gamma_0} \subset N$.*

Proof. We give an argument for Σ^+_0—the generalization to Σ^+_{0,λ_0} is obvious.

We argue by contradiction; i.e., assume there is a neighborhood N_0 of $\Sigma^+_0 \cap ([0,\gamma_0] \times W^{1,p}(0,1))$ such that $(\lambda_k, u_k) \in \Upsilon^+_{\varepsilon_k,\gamma_0}$ and $(\lambda_k, u_k) \notin N_0$ for all $k \in \mathbb{N}$. But for some subsequence, then we have $\lambda_k \to \lambda \in [0,\gamma_0]$ and $u_k \to u \in B_{2M_0}$ in $W^{1,p}(0,1)$. The latter follows from the fact that the uniform estimates (5.1) now hold with the constant $M_0 = \max\{M(\lambda) : 0 \leq \lambda \leq \gamma_0\}$ in place of $M(\lambda)$, and from a repetition of the arguments in this section. In particular, $(\lambda, u) \in \Sigma^+_0 \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big)$, i.e., $(\lambda_k, u_k) \in N_0$ for all $k \geq k_0(N_0) \in \mathbb{N}$, which is a contradiction. □

In view of properties (5.29) and Lemma 5.6, the main theorem of [1, section 3] yields that $\Sigma^+_0 \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big)$ and $\Sigma^+_{0,\lambda_0} \cap \big([0,\gamma_0] \times W^{1,p}(0,1)\big)$ are each connected and compact. Since this construction holds for any $\gamma_0 > 0$, we may conclude the following.

THEOREM 5.7. *Σ^+_0 and $\Sigma^+_{0,\lambda_0} \subset \mathbb{R} \times W^{1,p}(0,1)$ are each continua having projection $[0,\infty)$ on the parameter axis.*

Finally, in keeping with the notation of section 3, we define

$$\Gamma^+_{0,\lambda_0} \equiv \Sigma^+_{0,\lambda_0} \cap ([\lambda_0 + 1, \infty) \times W^{1,p}(0,1)), \tag{5.30}$$

which is the portion of the limiting continuum containing solutions of the double-well problem.

6. Minimizing properties of solutions for dead loading. For any weak solution of (2.3) with $\varepsilon = 0$ on the limiting continuum, $(\lambda, u_*) \in \Sigma^+_0$ or Γ^+_{0,λ_0}, we have from section 4 that the strain $z_* = u'_*$ suffers (at most) a single jump at some location $x = a(\lambda)$, where the Maxwell condition is satisfied; cf. (5.22). As mentioned in section 5, the latter is a necessary condition for u_* to minimize (2.1) for $\varepsilon = 0$ (at a given value of λ). In fact, given the nonconvexity of $W(\nu)$, the minimization of the functional (2.1) $\varepsilon = 0$ depends delicately upon the behavior of the loading potential $u \mapsto B(\lambda, u, x)$ [7], [12]; the minimum need not be attained in general. Accordingly, we focus in this section on the important case of "dead loading," viz.,

$$B(\lambda, u, x) \equiv -b(\lambda, x)u, \tag{6.1}$$

where the loading, $b = -B_u$, is now independent of u but otherwise satisfies condition (3.5) (note that (3.9) is now automatic).

We first observe that for a given value of $\lambda > 0$, a limiting strain $z_*(\lambda) = \frac{d}{dx}u_*(\lambda)$ satisfies the algebraic equation (5.12), which can be solved analytically as follows. We first consider single-well potential problems, i.e., we impose (3.2). For a given stress function $\sigma = W'(\nu)$, let $0 < \nu_1 < \nu_2$ denote the strains corresponding to the local maximum, $\sigma_m = \sigma(\nu_1)$, and local minimum, $\sigma(\nu_2)$, and let σ_M denote the Maxwell stress, as defined by the "equal-area" rule (5.22), cf. Figure 6.1. Since $x \mapsto z_*(x)$ and $x \mapsto F(\lambda, x)$ are each monotone decreasing, we can characterize solutions of (5.12) in terms of the magnitude of $F(\lambda, 0)$. In particular, for $F(\lambda, 0)$ sufficiently large, there are solutions such that $z_*(\lambda)$ has a single jump, where (5.22) now determines the location of the jump, $0 < a(\lambda) < 1$; cf. Figure 6.1. A typical profile corresponding to Figure 6.1 is depicted in Figure 6.2.

Finally, if we let $\sigma^{-1}_{\pm}$ denote the unique inverses of σ on $[0, \nu_1]$, $[\nu_2, \infty)$, respectively, we can express solutions of (5.12) analytically and hence, characterize Σ^+_0.

PROPOSITION 6.1. *If* $0 < F(\lambda_*, 0) < \sigma_M$, *then at* $\lambda = \lambda_*$ *there is only one (lower) solution* $(\lambda_*, u_*) \in \Sigma^+_0$, *where* $z_* = u'_*$ *is given by*

$$z_*(x, \lambda_*) = \sigma^{-1}_+(F(\lambda_*, x)). \tag{6.2}$$

If $F(\lambda_*, 0) > \sigma_m$, *then again there is only one (jump) solution* $(\lambda_*, u_*) \in \Sigma^+_0$, *where* $z_* = u'_*$ *now suffers a jump discontinuity:*

$$z_*(x, \lambda_*) = \begin{cases} \sigma^{-1}_-(F(\lambda_*, x)), & 0 \le x \le a(\lambda_*), \\ \sigma^{-1}_+(F(\lambda_*, x)), & a(\lambda_*) < x \le 1. \end{cases} \tag{6.3}$$

Here $0 < a(\lambda_*) < 1$ *is the unique solution of*

$$F(\lambda_*, a) = \sigma_M. \tag{6.4}$$

If $F(\lambda_*, 0) = \sigma_M$, *the lower solution* (6.2) *and the jump solution* (6.3), (6.4) *coincide (a.e. on* $[0, 1]$*). Consequently, there is only one solution point* $(\lambda_*, u_*) \in \Sigma^+_0$ *at* $\lambda = \lambda_*$, *where* u'_* *is characterized (say) by* (6.2) *a.e. on* $[0, 1]$. *Within a loading regime*

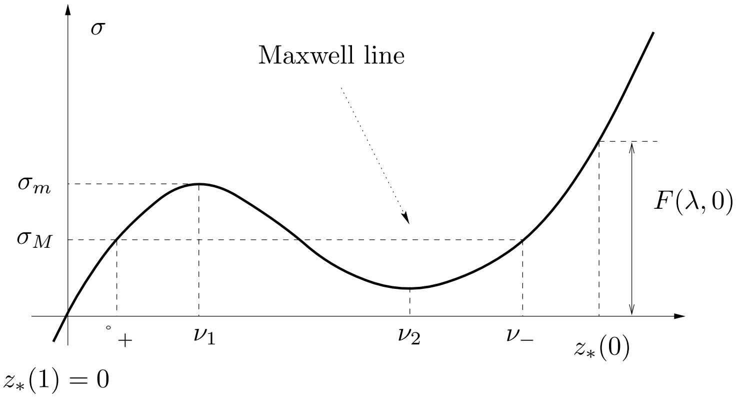


Fig. 6.1.

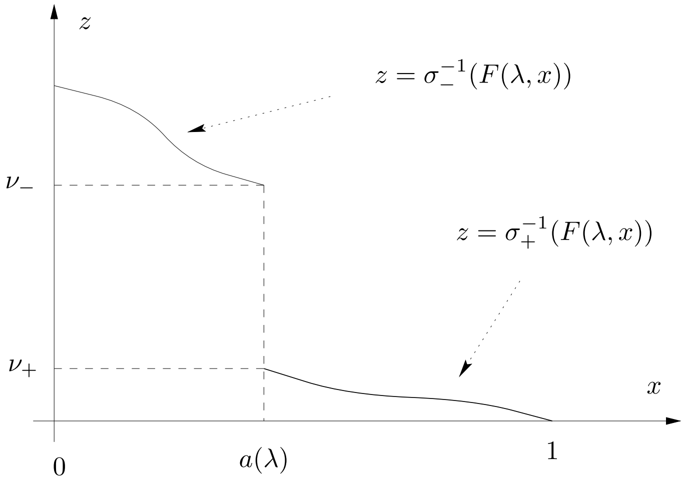


Fig. 6.2.

$\sigma_M < F(\lambda_*, 0) \leq \sigma_m$, *either or both solutions corresponding to* (6.2) *and* (6.3), (6.4), *respectively, may belong to* Σ_0^+. *Moreover, in that case suppose that* λ_* *belongs to some interval "I" such that* $F(\lambda, 0) \geq \sigma_M$ *for all* $\lambda \in I$ *and* $F(\lambda_0, 0) > \sigma_m$ *for some* $\lambda_0 \in I$. *Then* Σ_0^+ *contains a jump solution* (λ_*, u_*); *i.e.,* $u'_* = z_*$ *is characterized by* (6.3), (6.4).

Proof. The existence of the solutions (6.2) and (6.3), (6.4) to (5.12) and their uniqueness for $0 < F(\lambda_*, 0) < \sigma_M$ and $F(\lambda_*, 0) > \sigma_m$, respectively, are obvious. Accordingly, the associated solutions (λ_*, u_*) necessarily belong to the unbounded solution branch Σ_0^+ in those two cases. Within the regime $\sigma_M < F(\lambda_*, 0) \leq \sigma_m$ it is easy to see from (6.4) that $a(\lambda_*) > 0$, and the two solutions (6.2) and (6.3) of (5.12) are obviously distinct. Our limit analysis of section 5 provides no information about which one (or both) belongs to Σ_0^+ at $\lambda = \lambda_*$. However, given the extra loading information provided in the statement of the theorem, then (6.3), (6.4) define a curve of solutions for all $\lambda \in I$, and the jump solution at $\lambda = \lambda_0$ belongs to Σ_0^+ (by the same reasoning given at the beginning of the proof). Since Σ_0^+ is a continuum, it also contains the entire curve of jump solutions for all $\lambda \in I$, which includes λ_*.

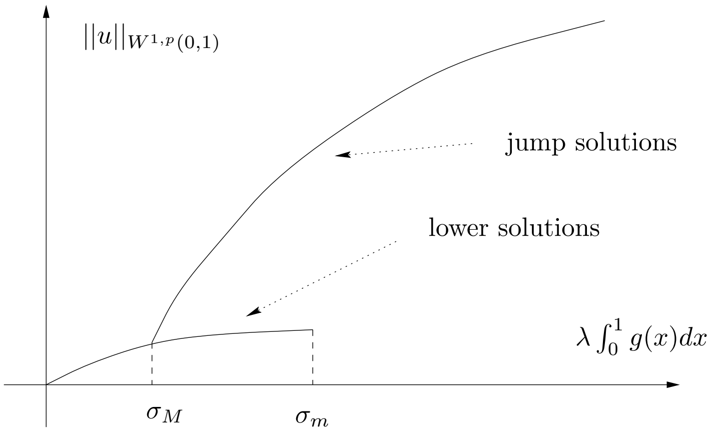


FIG. 6.3. *Solution branches for* (5.12) *in the case of a single-well potential with monotone loading.*

Finally, from the smoothness of $F(\lambda, x)$ and the monotonicity of $x \mapsto F(\lambda, x)$, we see that $a(\lambda)$ is continuous on any interval $[\lambda_1, \lambda_2]$ such that $F(\lambda, 0) \geq \sigma_M$ for all $\lambda \in [\lambda_1, \lambda_2]$. Moreover, $a(\lambda) \searrow 0$ as $\lambda \to \lambda_*$ whenever $F(\lambda_*, 0) = \sigma_M$ for $\lambda_* \in [\lambda_1, \lambda_2]$. Whence, (6.2) and (6.3), (6.4) are equal a.e. on $[0, 1]$ at $\lambda = \lambda_*$. □

In the important, special case of *monotone loading*, i.e., $b(\lambda, x) \equiv \lambda g(x)$, the loading condition stated at the end of Proposition 6.1 is always fulfilled. Moreover, (5.12) admits two intersecting curves of solutions: a terminating segment of lower solutions and an unbounded curve of jump solutions, both of which we depict schematically in Figure 6.3. Since Σ_0^+ is a continuum we immediately conclude the following.

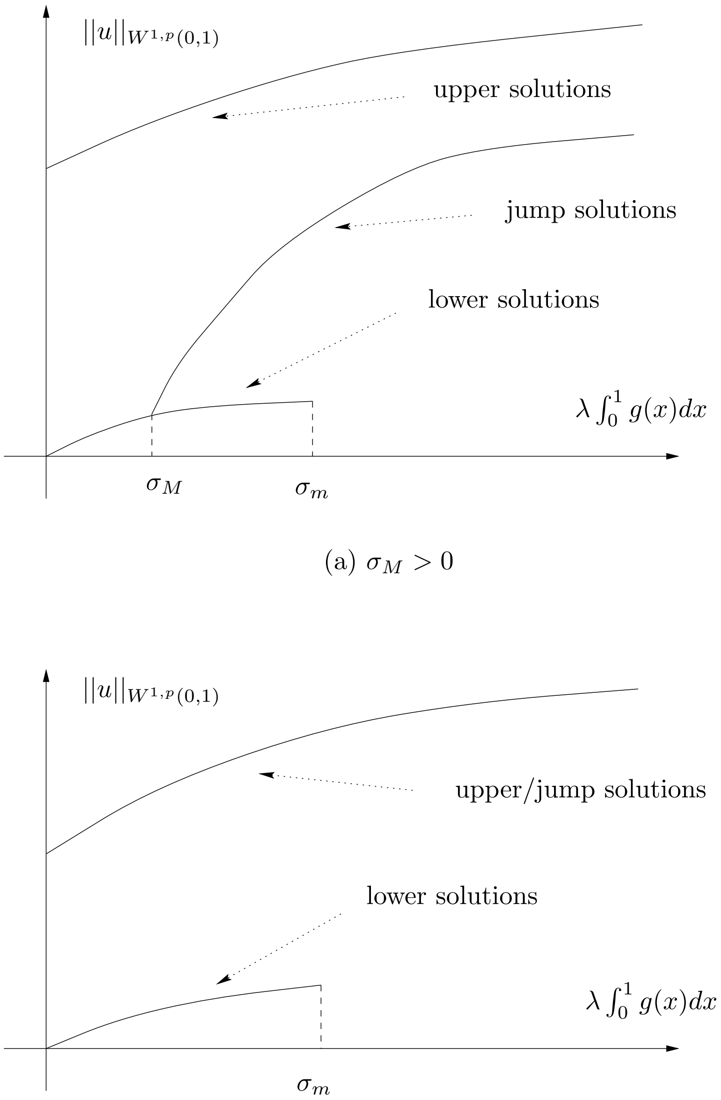
COROLLARY 6.2. *In the case of monotone loading, the limiting continuum* Σ_0^+ *contains a segment of lower solutions, characterized by* (6.2) *for* $0 \leq \lambda_* \int_0^1 g(x)dx \leq \sigma_M$, *and an unbounded curve of jump solutions; cf.* (6.3), (6.4) *for* $\lambda_* \int_0^1 g(x)dx \geq \sigma_M$. *In particular, for* $\sigma_M < \lambda_* \int_0^1 g(x)dx \leq \sigma_m$, Σ_0^+ *contains a jump solution* (λ_*, u_*).

Next we turn to two-well-potential problems. In this case, (6.2) and (6.3), (6.4) yield solutions of (5.12) (depending upon the magnitude of $F(\lambda_*, 0)$ as before). In addition, (5.12) now possesses another curve of smooth *upper* solutions,

$$\tilde{z}(x, \lambda) = \sigma_-^{-1}(F(\lambda, x)) \tag{6.5}$$

for all $\lambda \geq 0$. Note that $\tilde{z}(1, \lambda) = \nu_3$, where $\nu_3 > 0$ denotes the third zero of $\sigma(\nu)$; cf. Figure 4.1. Actually (6.3), (6.4), and (6.5) are distinct, provided that the Maxwell stress is positive, i.e., $\sigma_M > 0$. Our assumptions (2.6) allow for the possibility $\sigma_M = 0$, in which case $a(\lambda) \equiv 1$ (cf. (6.4)), and hence (6.3), (6.4), and (6.5) are identical on $0 \leq x < 1$. At any rate, for any value of $\lambda > 0$, (5.12) has either two or three solutions. This is best illustrated for monotone loading, $b(\lambda, x) = \lambda g(x)$, as shown schematically in Figure 6.4. However, unlike the single-well problem it is not immediately clear which solutions belong to the limiting set Γ_{0,λ_0}^+, cf. (5.30). We now take this up.

By construction, the unbounded continuum Σ_{0,λ_0}^+ (cf. Theorem 5.7) contains weak solutions of (2.3), $\varepsilon = 0$, for the single-well problem, for the homotopy between the single-well and the double-well problem, and for the double-well problem (the latter corresponding to Γ_{0,λ_0}^+); cf. sections 4 and 5. In particular, for the homotopy portion there is, by continuity, a number τ_0 with $0 < \tau_0 < 1$, or equivalently a parameter value θ_0 with $\lambda_0 < \theta_0 < \lambda_0 + 1$ (cf. (4.1) (4.4)), such that $\tilde{W}_\nu(\nu, \tau_0) = \tilde{\sigma}(\nu, \tau_0)$ has a double



(a) $\sigma_M > 0$

(b) $\sigma_M = 0$

FIG. 6.4. *Solution branches for* (5.12) *in the case of a double-well potential with monotone loading.*

zero at $\tau = \tau_0$, i.e., $\tilde{\sigma}(\nu_3, \tau_0) = \tilde{\sigma}_\nu(\nu_3, \tau_0) = 0$; cf. Figures 3.1 and 4.1. Moreover, all solution pairs belonging to $\Sigma^+_{0,\lambda_0} \cap ([0, \theta_0) \times W^{1,p}(0, 1))$ solve one-well problems; as such, they are characterized precisely as in Proposition 6.1. On the other hand, $\Sigma^+_{0,\lambda_0} \cap ((\theta_0, \infty) \times W^{1,p}(0, 1))$ contains only solutions of two-well problems.

PROPOSITION 6.3. *First suppose $\sigma_M > 0$. Then for any $\lambda_0 > 0$ and for all $\lambda_* \geq \lambda_0$, the limiting solution set Γ^+_{0,λ_0} contains only solution pairs (λ_*, u_*), where $z_* = u'_*$ is of the type* (6.2) *and/or* (6.3), (6.4)*, as described (in terms of the magnitude of $F(\lambda_*, 0)$) in Proposition* 6.1 *and in Corollary* 6.2 *for monotone loading. If $\sigma_M = 0$, then Γ^+_{0,λ_0} may contain upper solutions* (6.5) *(a.e. on* $[0, 1]$*) and/or lower solutions*

(6.2). *However, if* $F(\lambda_*, 0) > \sigma_m$ *for some* $\lambda_* \geq \lambda_0$ *(which is always satisfied, e.g., in the case of monotone loading) then* Γ^+_{0,λ_0} *contains the upper solution branch defined by* (6.5) *(a.e. on* $[0,1]$*) for all* $\lambda \geq \lambda_0$.

Proof. Let $\tilde{\sigma}_M(\tau)$ and $\tilde{\sigma}_m(\tau)$ denote the Maxwell stress and the local maximum stress, respectively, associated with (4.1) for each $0 \leq \tau \leq 1$. First, for $\sigma_M = \tilde{\sigma}_M(1) > 0$, we observe from (6.2)–(6.5) that

$$\|\tilde{z}(\lambda) - z_*(\lambda)\|_{L^p(0,1)} = \|\tilde{z}(\lambda) - z_*(\lambda)\|_{L^p(a(\lambda),1)} > 0 \tag{6.6}$$

for all $\lambda > 0$ (with the understanding that $a(\lambda) \equiv 0$ in (6.2)). In particular, the corresponding displacements, $\tilde{u}(\lambda)$ and $u_*(\lambda)$ (cf. (5.6)), satisfy

$$\|\tilde{u}(\lambda) - u_*(\lambda)\|_{W^{1,p}(0,1)} > 0 \text{ for all } \lambda > 0. \tag{6.7}$$

We now argue by contradiction, viz., suppose that Γ^+_{0,λ_0} and thus Σ^+_{0,λ_0} (cf. (5.30)) contain solution points $(\lambda_*, \tilde{u})$ where $\tilde{u}'$ is of the type (6.5) for $\lambda_* \geq \lambda_0$. We denote by "S" the curve of upper solutions defined by (6.5) for all $\tau_0 \leq \tau \leq 1$ and for all $\lambda \geq \lambda_0$. Recall that the solution set $R \equiv \Sigma^+_{0,\lambda_0} \cap ([0, \theta_0) \times W^{1,p}(0,1))$ contains no upper solutions. By virtue of (6.7), we then conclude that $\overline{R}$ and S are separated in $\mathbb{R} \times W^{1,p}(0,1))$. Since Σ^+_{0,λ_0} is a continuum, $\Sigma^+_{0,\lambda_0} \cap ([\theta_0, \infty) \times W^{1,p}(0,1))$ must contain lower and/or jump solutions, as described in Proposition 6.1. Again by (6.7), we then conclude that S is separated from Σ^+_{0,λ_0}, and hence $S \cap \Gamma^+_{0,\lambda_0} = \emptyset$; cf. (5.30).

Next we assume that $\sigma_M = \tilde{\sigma}_M(1) = 0$, in which case we have already noted that (6.5) and (6.3), (6.4) coincide a.e. on $[0,1]$. Thus, (6.2) and (6.5) define the only possible solutions of (5.12). In addition, if $F(\lambda_*, 0) > \sigma_m = \tilde{\sigma}_m(1)$ for some $\lambda_* \geq \lambda_0$, then (6.5) is the only solution of (5.12) at $\lambda = \lambda_*$. Again, in view of (6.7) and the fact that Σ^+_{0,λ_0} is a continuum, we conclude that the entire curve of upper solutions, (6.5) for all $\lambda \geq \lambda_0$, is contained in Σ^+_{0,λ_0} and hence in Γ^+_{0,λ_0} as well. □

To establish our "stability" (minimum) result, we need the following reasonable growth condition on the stored energy function:

$$|W'(\xi)| \leq c_1 + c_2|\xi|^{p-1} \tag{6.8}$$

for positive constants c_1, c_2, and integer $p \geq 1$.

Next we consider $(\lambda_*, u_*) \in \Sigma^+_0$ for single-well problems and $(\lambda_*, u_*) \in \Gamma^+_{0,\lambda_0}$ for double-well problems. In either case, for $\sigma_M > 0$, we henceforth assume that u_* is a lower solution (6.2) for $0 \leq F(\lambda_*, 0) \leq \sigma_M$ and a jump solution for $F(\lambda_*, 0) > \sigma_M$. For $\sigma_M = 0$, we assume that u_* is an upper solution (6.5) (a.e. on $[0,1]$). Loading conditions sufficient for these assumptions are given in Propositions 6.1 and 6.3. In particular, all such loading conditions are satisfied in the case of monotone loading.

THEOREM 6.4. *With* (λ_*, u_*) *as described above, suppose that* (6.8) *holds. Then* $V_0(\lambda_*, u)$ *attains its minimum at* u_*, *i.e.,*

$$\min_u \{V_0(\lambda_*, u) : u \in W^{1,p}(0,1), u(0) = 0\} = V_0(\lambda_*, u_*), \tag{6.9}$$

where V_0 *corresponds to* (2.1) *with* $\varepsilon = 0$, u_* *is the limiting displacement given by* (5.23)*, and* $1 \leq p < \infty$.

Proof. We first consider the convex minimization problem

$$\min_u \{V_0^{**}(\lambda_*, u) : u \in W^{1,p}(0,1), u(0) = 0\}, \tag{6.10}$$

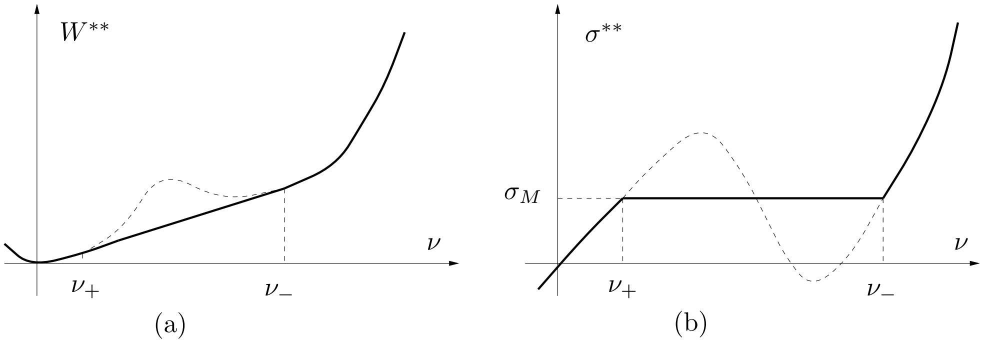


FIG. 6.5.

where $V_0^{**}(\lambda_*, u) = \int_0^1 [W^{**}(u') - b(\lambda_*)u]dx$, and

$$W^{**}(\xi) \equiv \sup\{g(\xi) : g(\xi) \leq W(\xi), g \text{ convex}\}. \tag{6.11}$$

In particular, we have (cf., e.g., [11, section A.2])

$$\sigma^{**}(\xi) = {W^{**}}'(\xi) = \begin{cases} \sigma(\xi), & \xi \in (-\infty, \nu_+] \cup [\nu_-, \infty), \\ \sigma_M, & \xi \in (\nu_+, \nu_-), \end{cases} \tag{6.12}$$

as depicted in Figure 6.5, where ν_+, ν_- and σ_M are as defined before. Condition (6.8) ensures that V_0^{**} is Gateaux differentiable; cf. [11, section 3.4].

Moreover, in view of (6.12) and the stated properties of u_*, we conclude that

$$\begin{aligned} < \delta V_0^{**}(\lambda_*, u_*), \eta > \; &\equiv \int_0^1 [\sigma^{**}(z_*)\eta' - b(\lambda_*)\eta]dx \\ &= \int_0^1 [\sigma(z_*)\eta' - b(\lambda_*)\eta]dx = 0, \end{aligned} \tag{6.13}$$

for all test functions $\eta \in W^{1,p}(0,1)$ satisfying $\eta(0) = 0$ (where the left-hand side of (6.13) denotes the Gateaux derivative of $u \mapsto V_0^{**}(\lambda_*, u)$ at u_*). In particular, u_* is a weak solution of the Euler–Lagrange equation for $V_0^{**}(\lambda_*, u)$. Also, by virtue of (2.5), (2.6), (6.10), and (6.11), we readily see that $u \mapsto V_0^{**}(\lambda_*, u)$ is convex on $\{u \in W^{1,p}(0,1) : u(0) = 0\}$. Hence, we conclude (cf. [11, section 3.1])

$$\begin{aligned} \min_u V_0^{**}(\lambda_*, u) &= V_0^{**}(\lambda_*, u_*) \\ &= V_0(\lambda_*, u_*), \end{aligned} \tag{6.14}$$

where the second equality follows from (6.10), (6.11) and the stated properties of u_*. Finally, from (6.11) we observe that

$$V_0^{**}(\lambda_*, u) \leq V_0(\lambda_*, u) \quad \text{for all } u \in W^{1,p}(0,1), \tag{6.15}$$

which, when combined with (6.14), gives the desired result. □

7. Concluding remarks. For dead loading (6.1) and $\varepsilon > 0$, the energy functional (2.1) is equivalent to

$$\tilde{V}_\varepsilon(\lambda, z) = \int_0^1 \left[\frac{\varepsilon}{2}(z')^2 + W(z) - F(\lambda, x)z\right] dx, \tag{7.1}$$

the direct minimization of which can be carried out along the lines of [3]. In particular, it can be shown that the global minimizer, say, $\tilde{z}_\varepsilon$, is monotone decreasing on $[0, 1]$, and moreover there is a subsequence $z_{\varepsilon_k} \to z_*$ pointwise on $[0, 1]$. This strongly suggests that (for each fixed λ) our solutions z_ε, obtained on global solution continua in sections 2–4, coincide with the minimizer $\tilde{z}_\varepsilon$. However, this need not be the case—e.g., the solution continuum $\mathcal{C}_\varepsilon^+$ (cf. section 3) could certainly contain turning points. Hence, for a given value of λ, there can be three (or more) distinct, monotone decreasing solutions, at least one of which must be a saddle point of (7.1). However, since both (sub)sequences converge to z_*, it is straightforward to show that

$$\|z_{\varepsilon_k} - \tilde{z}_{\varepsilon_k}\|_{L^p(0,1)} \to 0 \quad \text{as } k \to \infty. \tag{7.2}$$

We emphasize that the results of sections 3–5 yield global solution continua, in the limit of vanishing capillarity, for a general class of parameter-dependent live loadings. For sufficiently large loading (e.g., monotonic in the parameter), we prove that the limiting solution continuum possesses weak solutions characterized by a single jump, at which the two Erdmann–Weierstrass corner conditions [14] are always satisfied. The specialization to dead loading in section 6 represents the simplest case for which satisfaction of the corner conditions is enough to deduce that the limiting weak solution also minimizes the (zero-capillarity) energy. Of course in the dead-load case, the direct minimization of energy is also fruitful. Note that the general results of section 5 hold in cases where the direct minimization of energy (zero-capillarity) may be difficult or even ill-posed; e.g., cf. [12]. Also, the minimum energy approach provides no information about solution continua as the loading is varied.

Generalizations of our results, e.g., to more general loadings and/or higher-dimensional models, are complicated by the need for uniform a priori bounds (e.g., (5.1)) for convergence. Recently, a similar analysis of a two-dimensional Cahn–Hilliard problem with rectangular symmetry has been analyzed by the second-named author [20]. This is essentially a two-dimensional version of the problem considered in [8]. In [20] the a priori estimates rely upon a subtle combination of the maximum principle and symmetry arguments related to those in [16]. Also, a two-dimensional version of Helly's pointwise convergence theorem for functions having appropriate monotonicities is obtained there. Using the Erdmann–Weierstrass corner conditions, the minimization properties of the limiting weak solutions are proved in [21].

Finally we note that global branches of weak solutions for some elastic string problems have been obtained by regularization and singular limits in [2]. There the possible vanishing of the tension causes the singularity, and the regularization chosen is not physical. Here our small-capillarity solutions are also of physical significance.

Acknowledgment. The authors thank Phoebus Rosakis for valuable comments throughout various stages of this work.

REFERENCES

[1] J.C. Alexander, *A primer on connectivity,* in Fixed Point Theory, E. Fadell and G. Fournier, eds., Springer-Verlag, New York, 1981, pp. 455–483.

[2] J.C. Alexander, S.S. Antman, and S.-T. Deng, *Nonlinear eigenvalue problems for the whirling of heavy elastic strings*, II: *New methods of global bifurcation theory,* Proc. Roy. Soc. Edinburgh Sect. A, 93 (1983), pp. 197–227.

[3] N.D. Alikakos and K.C. Shaing, *On the singular limit for a class of problems modelling phase transitions*, SIAM J. Math. Anal., 18 (1987), pp. 1453–1462.

[4] G. Aubert and R. Tahraoui, *Théoréms d'existence pour des problémes du calcul des variationes du type: $Inf \int_0^L f(x, u'(x))dx$ et $Inf \int_0^L f(x, u(x), u'(x))dx$.*, J. Differential Equations, 33 (1979), pp. 1–15.

[5] J. Ball and R.D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1992), pp. 13–52.

[6] M.A. Biot, *Mechanics of Incremental Deformations*, John Wiley, New York, 1965.

[7] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincairé Anal. Non Linéaire, 7 (1990), pp. 97–106.

[8] J. Carr, M.E. Gurtin, and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Rational Mech. Anal., 86 (1984), pp. 317–351.

[9] J.W. Cahn and J.E. Hilliard, *Free energy of a nonuniform system* I. *Interfacial free energy*, J. Chem. Physics, 28 (1958), pp. 258–267.

[10] M. Crandall and P.H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[11] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, New York, 1989.

[12] I. Ekeland, *Discontinuité de champs hamiltoniens et existence de solutions optimales en calcul des variations*, Inst. Hautes Études Sci. Publ. Math., 47 (1977), pp. 5–32.

[13] J.L. Ericksen, *Equilibrium of bars*, J. Elasticity, 5 (1975), pp. 191–202.

[14] G.M. Ewing, *Calculus of Variations*, Dover, New York, 1985.

[15] T.J. Healey and H. Kielhöfer, *Symmetry and nodal properties in the global bifurcation analysis of quasi-linear elliptic equations*, Arch. Rational Mech. Anal., 113 (1991), pp. 299–311.

[16] T.J. Healey and H. Kielhöfer, *Preservation of nodal structure on global bifurcating solution branches of elliptic equations with symmetry*, J. Differential Equations, 106 (1993), pp. 70–89.

[17] T.J. Healey and P. Rosakis, *Unbounded branches of classical injective solutions to the forced displacement problem in nonlinear elastostatics*, J. Elasticity, 49 (1997), pp. 65–78.

[18] T.J. Healey and H. C. Simpson, *Global continuation in nonlinear elasticity*, Arch. Rational Mech. Anal., (143) 1998, pp. 1–28.

[19] H. Kielhöfer, *Multiple eigenvalue bifurcation for Fredholm operators*, J. Reine Angew. Math., 358 (1985), pp. 104–124.

[20] H. Kielhöfer, *Pattern formation of the stationary Cahn-Hilliard Model*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1219–1243.

[21] H. Kielhöfer, *Minimizing sequences selected via singular perturbations and their pattern formation*, Arch. Rational Mech. Anal., to appear.

[22] R. Kohn and S. Müller, *Surface energy and microstructures in coherent phase transitions*, Comm. Pure Appl. Math., 47 (1994), pp. 405–435.

[23] A.E. Lifshitz and G.L. Rybnikov, *Dissipative structures and Couette flow in a non-Newtonian fluid*, Sov. Phys. Dokl., 30(4) (1985), pp. 276–278.

[24] L.A. Lusternik and V.J. Sobolev, *Elements of Functional Analysis*, Gordon and Breach, New York, 1968.

[25] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.

[26] A. Novick-Cohen and L.A. Segel, *Nonlinear aspects of the Cahn-Hilliard equation*, Phys. D, 10 (1984), pp. 277–298.

[27] R.W. Ogden, *Non-linear elastic deformations*, Ellis Horwood Ltd., Chichester, UK, 1984.

[28] P.H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[29] P. Rosakis, *Compact zones of shear transformation in an anisotropic solid*, J. Elasticity, 40 (1992), pp. 1163–1195.

[30] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.

[31] N. Triantafyllidis and E. Aifantis, *A gradient approach to localization of deformation* I. *Hyperelastic materials*, J. Elasticity, 16 (1986), pp. 225–237.

[32] N. Triantafyllidis and S. Bardenhagen, *On higher gradient continuum theories in* 1-*D nonlinear elasticity. Derivation from and comparison to the corresponding discrete models*, J. Elasticity, 33 (1993), pp. 259–293.

[33] J.D. Van Der Waals, *The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density (in Dutch)*, Verhandel. Konink. Akad. Weten. Amsterdam (sect. 1), 1(8) (1893).

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1332–1350

A COMPLETE SOLUTION CHARACTERIZING SMOOTH REFINABLE FUNCTIONS[*]

VLADIMIR PROTASOV[†]

Dedicated to the memory of Mary Panfyorova

Abstract. In this paper we study univariate two-scale refinement equations $\varphi(x) = \sum_{k\in\mathbb{Z}} c_k \varphi (2x-k)$, where the coefficients $c_k \in \mathbb{C}$ satisfy an exponential decay assumption. We show that any refinement equation that has a smooth solution can be reduced to the well-studied case of complete sum rules: $\sum_k (-1)^k k^n c_k = 0, \quad n = 0, \dots, L$, where L depends on regularity of the solution. This result makes it possible to extend previously known results on refinable functions and subdivision schemes from the case of complete sum rules to the general case. As a corollary we obtain sharp necessary conditions for the existence of smooth refinable functions and the convergence of corresponding cascade algorithms. Other applications concern polynomial spaces spanned by integer translates of a refinable function and one special property of linear operators associated to refinement equations.

Key words. refinement equations, regularity, reducibility, tree, cascade algorithm

AMS subject classifications. 26C10, 39B32, 42A05, 42A38

PII. S0036141098342969

1. Introduction.

Functional equations of the type

$$\varphi(x) = \sum_{k=0}^{N} c_k \varphi(qx-k)$$

(*q-scale refinement equations*), where $c_k \in \mathbb{C}$, $q \geq 2$ is an integer, have found a lot of applications in wavelets theory [C], [CD1], [DL2], [CH], [HC], in subdivision algorithms in approximation theory and curve design [CDM], [DGL1], [DGL2], [DD], [M], [DyL], and in probability theory [DDL]. We restrict ourselves in this paper to the case $q = 2$, especially as this case is the most important one in applications. Nevertheless all our results can be applied for general integer values of $q \geq 2$. We say that a function $\varphi \in L_1(\mathbb{R})$ is a *refinable function* if it has a compact support and satisfies a two-scale refinement equation

$$\varphi(x) = \sum_{k=0}^{N} c_k \varphi(2x-k). \tag{1}$$

It was shown in [DL1] that any refinement equation has at most one compactly supported L_1-solution up to normalization. This solution, if it exists at all, has its support in the segment $[0, N]$ [DD], [DL1]. A necessary condition for the existence of this solution is $\sum c_k = 2^r$ for some natural r. Moreover, it is sufficient to consider only the main case

$$\sum_k c_k = 2, \quad \int \varphi(x)dx = 1, \tag{2}$$

[*] Received by the editors August 4, 1998; accepted for publication (in revised form) May 17, 1999; published electronically May 17, 2000.
http://www.siam.org/journals/sima/31-6/34296.html

[†] Department of Economics, Erasmus Universitet of Rotterdam, H-17-19, Burgemeester Oudlaan-50, Rotterdam, 3062 PA, The Netherlands (protassov@few.eur.nl).

because the other cases can be reduced to (2) by integration of the refinable function and a suitable normalization [DL1]. Thus we consider (1) and assume below that condition (2) is satisfied.

The existence and regularity of refinable functions have been studied by many authors in great detail [DL1], [DL2], [CH], [HC], [CDM], [LW], [V].

In much of the literature this problem is investigated by the frequency domain approach, i.e., by considering the Fourier transform of (1)

$$\widehat{\varphi}(\xi) = m_0\Big(\frac{\xi}{2}\Big)\widehat{\varphi}\Big(\frac{\xi}{2}\Big), \tag{3}$$

where the trigonometric polynomial $m_0(\xi) = \frac{1}{2}\sum_{k=0}^{N} c_k e^{-i\xi k}$ is the *mask* of the equation. Another approach, which uses linear algebraic tools in the time domain, has also been successfully applied [DL2], [CH], [CDM], [W1], [W2]. This approach is based on the fact that values of a continuous refinable function $\varphi(x)$ can be calculated consequently at all dyadic points. An explicit formula for $\varphi(x)$ can be expressed by infinite products of the two special matrices T_0 and T_1 (see [DL1], [CH] or formulas (21) and (22) in this paper). In the case $\varphi \in L_p$ the calculation of $\varphi(x)$ at dyadic points is replaced by the calculation of integrals of φ over the dyadic intervals (see [LW], [W2], or formula (23)). The existence and smoothness of solutions are studied by properties of corresponding linear operators (see [DL1] for details).

Many of the works on refinable functions deal with the one special case (the case of *complete sum rules*). Namely, if we study L times differentiable solutions of (1), then we assume that the coefficients $\{c_k\}$ satisfy the *sum rules* of order L, i.e., the following $L+1$ conditions:

$$m_0^{(n)}(\pi) = \frac{1}{2}\sum_k (-1)^k k^n c_k = 0, \quad n = 0, \dots, L. \tag{4}$$

In particular the first sum rule

$$\sum_k c_{2k} = \sum_k c_{2k+1} = 1 \tag{5}$$

has been assumed in almost all papers devoted to continuous or L_1-solutions of (1). The sum rules, which were elaborated in [DL1] and [CDM], simplify many problems on refinable functions. In the frequency domain approach the sum rules make it possible to reduce the mask of the equation to the form

$$m_0(\xi) = \left(\frac{1+e^{-i\xi}}{2}\right)^{L+1} q(\xi)$$

and then analyze the smoothness of refinable function by estimating the polynomial $q(\xi)$. This method was applied in [DD], [D]. In the time domain approach the sum rules allow us to obtain very explicit expressions for the common eigenspaces of operators T_0 and T_1 and actually to reduce these operators to a block-diagonal form [DL2], [HC].

A quite different approach to the study of refinement equations was proposed in the works of Herve [He1], [He2] and of Cohen and Daubechies [CD2]. These works deal with refinement equations with infinitely many terms

$$\varphi(x) = \sum_{k\in\mathbb{Z}} c_k \varphi(2x-k) \tag{6}$$

under the exponential decay assumption on the coefficients $\{c_k\}$:

$$|c_k| < Ce^{-\gamma|k|}, \quad k \in \mathbb{Z}, \tag{7}$$

for some constants $\gamma > 0,\ C > 0$. The technique based on the study of spectral properties of the *transfer operator* made it possible to obtain very sharp information for the Sobolev regularity of solution of (6). These results were also obtained for the case of holding sum rules: $m_0^{(n)}(\pi) = 0,\ n = 0,\dots,L$, where $m_0(\xi) = \frac{1}{2}\sum_k e^{-i\xi k}$ is the mask of the equation. The estimation of the Sobolev exponent established in [He1], [He2], and [CD2] actually depends on the order L of sum rules.

In addition, a lot of interesting results on local regularity [DL2], fractal properties [W2], [DL2], and L^2-regularity [LW], [CD1] of refinable functions were obtained for the case when sum rules (4) hold. However, this is just a particular, although very important, case of refinement equations. The sum rules are not necessary for the existence of smooth refinable functions. Corresponding examples are well known [DL2], [CDM]. For instance, the equation

$$\varphi(x) = \frac{1}{3}\varphi(2x) + \frac{2}{3}\varphi(2x-1) + \frac{1}{3}\varphi(2x-2) + \frac{2}{3}\varphi(2x-3) \tag{8}$$

has a continuous compactly supported solution (see [P]); nevertheless, none of the sum rules holds. Moreover, as we shall see in section 4, there are a lot of refinement equations without the sum rules that have, nevertheless, “good” smooth solutions. However, all attempts to get rid of the sum rules and consider the general case have always significantly complicated the problem. The estimations of regularity of refinable functions obtained in the works [D], [DL2], [DD], [He2], [Hu], [CD2] depend on the order of sum rules that hold for corresponding refinement equations. As a rule the results obtained for the case of complete sum rules are much sharper than the results in the general case. Now let us formulate the following questions.

Question 1. Is the case of complete sum rules a very special case in the theory of refinement equations? Is it possible to extend the results mentioned above from this case to all refinable functions?

Question 2. Why do all the most interesting examples of refinement equations which have ever appeared in applications satisfy the sum rules? Can refinement equations without sum rules be useful in applications?

Question 3. How to obtain necessary conditions for the existence of smooth solutions of refinement equations that would be sharp in some sense?

In the next section we formulate the “theorems of reduction.” Then we discuss each of these questions. Questions 2 and 3 have been considered in the literature, and some important results were obtained. We will mention them in the next section.

2. Theorems of reduction. Denote, as usual, by $C^L(\mathbb{R}),\ L \geq 0$ the space of L times continuously differentiable functions. Let $C^0(\mathbb{R})$ denote the space $C(\mathbb{R})$; let $W_p^L(\mathbb{R}), L \geq 0, p \in [1,\infty]$ denote the Sobolev space (the space of functions whose Lth derivatives are in $L^p(\mathbb{R})$), and let $W_p^0(\mathbb{R})$ denote the space $L^p(\mathbb{R})$. Let us denote by Π_L the space of polynomials of degree at most L. For a compactly supported function $\varphi(x)$ denote by $\mathcal{V}_\varphi$ the space that is spanned by integer translates of the function φ, i.e.,

$$\mathcal{V}_\varphi = \left\{ \sum_{k\in\mathbb{Z}} \lambda_k \varphi(x-k), \quad \lambda_k \in \mathbb{C} \right\}.$$

Now consider an arbitrary function $f \in C(\mathbb{R})$. Let L be the maximal integer such that $f \in C^L(\mathbb{R})$. Then the Hölder exponent of this function is $L + \beta_m$, where β_m is the supremum of the values β such that for some constant $C(\beta)$ the inequality $\|f^{(L)}(x+y) - f^{(L)}(x)\|_{C(\mathbb{R})} \leq C(\beta)|y|^\beta$ holds for any $y \in \mathbb{R}$. The Hölder exponent in the space W_p^L is defined as above by the substitution of $L^p(\mathbb{R})$ instead of $C(\mathbb{R})$. The Sobolev exponent s_p of a function $\varphi(x)$ is defined for any $p \in [1, +\infty)$ as

$$s_p(\varphi) = \sup \left\{ s \in \mathbb{R} \Big| \int\limits_{\mathbb{R}} |\widehat{\varphi}(\xi)|^p (1 + |\xi|^{ps}) d\xi < \infty \right\}.$$

We say that two functions have equal smoothness if both the following conditions are satisfied:

(1) These functions belong to the same functional space (C^L or W_p^L) and have equal Hölder exponents in that space.

(2) They have equal Sobolev exponent for any $p \geq 1$.

We also say that (6) (under conditions (2) and (7)) and its L^1-solution $\varphi(x)$ are *reducible* if for some integer $L \geq 0$, $\varphi \in W_1^L(\mathbb{R})$ (or $s_p(\varphi) > L$ for some $p \geq 1$) hold, and the sum rules of order L do not hold, i.e., at least one of the $L+1$ conditions (4) is not satisfied.

The opposite case is said to be *irreducible*, or the *case of complete sum rules.* If we denote by $L_{\max}$ the maximal integer $L \geq 0$ such that $\varphi \in W_1^L(\mathbb{R})$ or $s_p(\varphi) > L$ for some $p \geq 1$, then we have the following:

A refinement equation and its L^1-solution are irreducible if and only if sum rules of order $L_{\max}$ hold.

Now we formulate the first theorem of reduction.

THEOREM 1 (the case of finitely many terms). *For any reducible solution $\varphi(x)$ of* (1) *there exist a constant $\theta \in \mathbb{R}$ and a function $\psi(x)$ such that the following hold:*

(1) $\theta = \frac{\pi(2l+1)}{2^n}$ *for some $n \in \{0, \dots, N-2\}$ and $l \in \{0, \dots, 2^n - 1\}$. Moreover, $\frac{\theta}{2}$ and $\frac{\theta}{2} + \pi$ are roots of the equation $m_0(\xi) = 0$.*

(2) $\varphi(x) = \psi(x) - e^{i\theta}\psi(x-1)$.

(3) $\operatorname{supp} \psi(x) \subset [0, N-1]$.

(4) *The function $\psi(x)$ satisfies the refinement equation*

$$\psi(x) = \sum_{k=0}^{N-1} \tilde{c}_k \psi(2x - k),$$

where $\tilde{c}_k$ are coefficients of the polynomial

$$\sum_{k=0}^{N-1} \tilde{c}_k e^{-i\xi k} = 2\tilde{m}_0(\xi) = 2\frac{m_0(\xi)(1 - e^{i(\theta-\xi)})}{1 - e^{i(\theta - 2\xi)}}.$$

(5) *The functions φ and ψ have an equal smoothness.*

(6) $\mathcal{V}_\varphi = \mathcal{V}_\psi$.

Thus it is possible to pass from every reducible equation to an equation with complete sum rules by at most $N-1$ steps of reduction. Each step can be realized easily by exhaustion of a finite number of values θ (item 1 of Theorem 1). Furthermore, this passage preserves the smoothness of function and decreases the order of equation. Therefore, (in relation to Question 1) the case of complete sum rules is not particular but rather general. The results about the existence and smoothness of refinable functions can be extended from the case of complete sum rules to all refinable functions without any difficulty.

To generalize this result to refinement equations with infinitely many terms we formulate the second theorem of reduction. We consider (6) and assume that there exist constants $\gamma, C > 0$ such that the fast decay condition (7) is satisfied. Also we assume as before conditions (2).

Theorem 2. *Suppose $\varphi(x)$ is an L^1-solution of a reducible refinement equation* (6)*; then there exist a constant $\theta \in \mathbb{R}$ and a function $\psi(x)$ such that the following hold:*

(1) $\theta = \frac{\pi(2l+1)}{2^n}$ *for some integers $n \geq 0$ and $l \in \{0, \dots, 2^n - 1\}$. Moreover, the numbers $\frac{\theta}{2}$ and $\frac{\theta}{2} + \pi$ are roots of the equation $m_0(\xi) = 0$.*

(2) $\varphi(x) = \psi(x) - e^{i\theta}\psi(x-1)$.

(3) *The function $\psi(x)$ satisfies the refinement equation*

$$\psi(x) = \sum_{k\in\mathbb{Z}} \tilde{c}_k \psi(2x - k),$$

where the coefficients $\tilde{c}_k$ are defined from the equality

$$\sum_{k\in\mathbb{Z}} \tilde{c}_k e^{-i\xi k} = 2\tilde{m}_0(\xi) = 2\frac{m_0(\xi)(1 - e^{i(\theta-\xi)})}{1 - e^{i(\theta-2\xi)}}.$$

(4) *The functions φ and ψ have an equal smoothness.*

Condition (7) implies that the mask m_0 of (6) is holomorphic in the strip $\Gamma = \{\xi \in \mathbb{C}, \ \ |Im\ \xi| < \gamma\}$. Therefore the mask has at most finitely many zeros on the segment $[0, 2\pi]$. Thus, it is possible to pass from every refinement equation having a smooth solution to an irreducible equation by finitely many steps of reduction. So the technique, which was developed in the works [He1], [He2], and [CD2] (to estimate the Sobolev regularity of solutions of (6) in terms of the spectral radius of the corresponding transfer operator), is extended from the case of complete sum rules to the general case. Let us also note that Theorem 2 can be formulated without the assumption $\varphi \in L^1(\mathbb{R})$ (Remark 1). On the other hand, the exponential decay assumption (condition (7)) is essential for Theorem 2 to hold (Remark 9).

Question 2 has been discussed in the literature. In particular it was shown in [DGL2] that the first sum rule (condition (5)) is necessary for the convergence of subdivision scheme or, in other terms, for convergence of the *cascade algorithm* to a continuous refinable function. In section 6 we shall obtain a sharp necessary condition for a cascade algorithm to converge to a C^L-refinable function (Corollaries 7 and 8). Another result of that section reduces the problem of convergence of cascade algorithms from the general case to the case of complete sum rules (Proposition 2).

The sum rules have also found applications in the study of wavelets. Daubechies in [D] proved that the sum rules of order L are necessary for the orthogonality of wavelets constructed by refinable functions from C^L. Then Villemoes [V] showed that the first sum rule is necessary for refinable functions to possess the Riesz basis property in L^2. In section 4 we supplement these results. Namely we prove that for any $p \in [1, \infty]$ the sum rules of order L are necessary for a refinable function from W_p^L to possess the Riesz basis property in L^p (Corollary 3). This means in particular that integer translates of a reducible function cannot form a multiresolution analysis (see [C] for definitions).

We next make several comments about Question 3. Some necessary conditions for the existence of a W_p^L-solution of refinement equation were discussed in [CH], [W1] (the condition in terms of the joint spectral radius), and in [CDM] (the condition in terms of the geometric mean). We shall obtain a condition on zeros of polynomial $m_0(\xi)$ (Corollary 1) that can be easily verified for a given refinement equation (Remark 3). This condition is sharp in the sense that all its cases are realized (Corollary 2).

The other results of this paper concern the polynomial space spanned by integer translates of refinable function (Theorem 3) and special properties of the linear operators T_0 and T_1 associated to a reducible refinement equation (section 5).

3. Proof of Theorems 1 and 2.

The proof will be split into several lemmas.

LEMMA 1. *Suppose a function $\varphi(x)$ is an L^1-solution of (6) such that $\varphi \in W_1^L(\mathbb{R})$ or $s_p(\varphi) > L$, where $L \geq 0,\ p \geq 1$; then*

$$\widehat{\varphi}^{(r)}(2\pi n) = 0 \tag{9}$$

for every $n \in \mathbb{Z} \setminus \{0\}$ and $r = 0, \ldots, L$.

Proof. It follows from condition (7) that the function $m_0(\xi)$ is holomorphic in the strip $\Gamma = \{\xi \in \mathbb{C},\ \ |Im\ \xi| < \gamma\}$. Further, by iterating equality (3) k times ($k \geq 1$) we obtain

$$\widehat{\varphi}(\xi) = \widehat{\varphi}\left(\frac{\xi}{2^k}\right)\prod_{l=1}^{k} m_0\left(\frac{\xi}{2^l}\right). \tag{10}$$

It follows from (2) that $m_0(0) = 1$, hence the product $\prod_{l=1}^{k} m_0(\frac{\xi}{2^l})$ converges uniformly in every compact subset of Γ as $k \to \infty$. This shows that the function $\widehat{\varphi}(\xi)$ is holomorphic in Γ. Now if we recall that $\widehat{\varphi}(0) = 1$ (condition (2)), we obtain

$$\widehat{\varphi}(\xi) = \prod_{l=1}^{\infty} m_0\left(\frac{\xi}{2^l}\right), \quad \xi \in \Gamma. \tag{11}$$

(This is a well-known formula of solutions of refinement equations; see [D].)

Further, take arbitrary $\varepsilon \in (0, \min\{\gamma, 2\pi\})$ such that $\widehat{\varphi}(\xi)$ does not vanish in the disk $\Omega_\varepsilon = \{\xi \in \mathbb{C},\ |\xi| < \varepsilon\}$. Take also a $\delta \in [0, \varepsilon)$ and integers $n \neq 0,\ k \geq 1$. Substituting $2^{k+1}\pi n + \delta$ for ξ in (10), we get

$$\widehat{\varphi}(2\pi n 2^k + \delta) = \widehat{\varphi}\left(2\pi n + \frac{\delta}{2^k}\right)\prod_{l=1}^{k} m_0\left(2\pi n 2^{k-l} + \frac{\delta}{2^l}\right) = \widehat{\varphi}\left(2\pi n + \frac{\delta}{2^k}\right)\prod_{l=1}^{k} m_0\left(\frac{\delta}{2^l}\right).$$

It follows from (10) that there is a constant $C_\varepsilon > 0$ such that the inequality

$$\inf_{z \in \Omega_\varepsilon}\left|\prod_{l=1}^{k} m_0(\frac{z}{2^l})\right| > C_\varepsilon$$

holds for all $k \geq 1$. Therefore,

$$|\widehat{\varphi}(2\pi n 2^k + \delta)| > C_\varepsilon |\widehat{\varphi}(2\pi n + \delta 2^{-k})|. \tag{12}$$

Now suppose that $\varphi(x)$ is in $W_1^L(\mathbb{R})$; then $\xi^L \widehat{\varphi}(\xi) \to 0$ as $\xi \to \infty$ along the real axis. Combining this and (12) we see that for any $\delta \in [0, \varepsilon)$

$$(2\pi n 2^k + \delta)^L \widehat{\varphi}\left(2\pi n + \frac{\delta}{2^k}\right) \to 0 \text{ as } k \to \infty.$$

Hence the function $\widehat{\varphi}(\xi)$ has zero of order at least $L+1$ at the point $\xi = 2\pi n$. This proves Lemma 1 in the case $\varphi \in W_1^L$.

In the case $s_p > L$ we have $\int |\widehat{\varphi}(\xi)|^p(1+|\xi|^{pL})|d\xi < \infty$. On the other hand,

$$\int |\widehat{\varphi}(\xi)|^p(1+|\xi|^{pL})|d\xi \geq \sum_{k\in\mathbb{N}} \int\limits_{\varepsilon/2}^{\varepsilon} |\widehat{\varphi}(2\pi n 2^k + \delta)|^p(1+|(2\pi n 2^k+\delta)|^{pL})d\delta$$

$$> \sum_{k\in\mathbb{N}} C_\varepsilon \int\limits_{\varepsilon/2}^{\varepsilon} |\widehat{\varphi}(2\pi n + \delta 2^{-k})|^p(1+|(2\pi n 2^k+\delta)|^{pL})d\delta.$$

Since this series converges, we see that

$$\int\limits_{\varepsilon/2}^{\varepsilon} |\widehat{\varphi}(2\pi n + \delta 2^{-k})|^p(1+|(2\pi n 2^k+\delta)|^{pL})d\delta \to 0 \quad \text{as } k \to \infty.$$

Applying the mean value theorem, we get

$$|\widehat{\varphi}(2\pi n + \eta_k 2^{-k})|^p(1+|(2\pi n 2^k+\eta_k)|^{pL}) \to 0 \quad \text{as } k \to \infty,$$

where $\eta_k \in (\varepsilon/2, \varepsilon)$. This implies that the function $\widehat{\varphi}(\xi)$ has zero of order at least $L+1$ at the point $\xi = 2\pi n$. Lemma 1 is proved. □

Before we formulate the next lemma let us introduce some notation.

Consider a binary tree. The vertex of the tree is said to be an *nth level vertex* if its distance to the root of the tree is equal to n. The root has level 0. To every vertex of the tree we shall associate a number as follows: put π at the root, then put $\frac{\pi}{2}$ and $\frac{3\pi}{2}$ at the vertices on the first level. Let the number α be associated to a vertex on the nth level; then the numbers $\alpha/2$ and $\alpha/2+\pi$ are associated to its neighbors on the $(n+1)$st level. Thus there are the numbers

$$\frac{\pi(2l+1)}{2^n}, \qquad l = 0, \ldots, 2^n - 1,$$

on the nth level of the tree. For convenience we shall identify a vertex and the corresponding number. We shall call a family of vertices (or numbers) $\mathcal{A}$ a *minimal cut set* (of the tree) if every infinite path (all the paths are without backtracking) starting at the root of the tree includes exactly one element of $\mathcal{A}$. For instance a one-element set $\mathcal{A} = \{root\} = \{\pi\}$ is a minimal cut set. It is easily shown that every minimal cut set is finite. We define the *type* of the set $\mathcal{A}$ to be the maximal level of its elements.

Definition 1. We say that (6) satisfies *the condition of order* L ($L \geq 0$) if there exist $L+1$ minimal cut sets of the tree $\mathcal{A}_0, \ldots, \mathcal{A}_L$, perhaps intersecting, such that the polynomial $\prod_{k=0}^{L}\prod_{\alpha\in\mathcal{A}_k}(1-e^{i(\alpha-\xi)})$ is a divisor of the mask $m_0(\xi)$.

In other words, all the sets $\mathcal{A}_0, \ldots, \mathcal{A}_L$ consist of roots of the mask $m_0(\xi)$ and if a vertex α belongs to k sets simultaneously ($k \geq 1$), then α is a root of $m(\xi)$ with multiplicity at least k.

It is easy to see that the condition of order L is equivalent to the following one: *every infinite path* $\alpha_0 \to \alpha_1 \to \cdots$ *starting at the root of the tree includes at least* $L+1$ *zeros of the function* $m_0(\xi)$ *(counting with multiplicity).*

LEMMA 2. *An L^1-solution $\varphi(\xi)$ of refinement equation* (6) *satisfies the equality* (9) *for every $n \in \mathbb{Z} \setminus \{0\}$ and $r = 0, \dots, L$ if and only if this equation satisfies the condition of order L.*

Proof. Suppose the function $\varphi(\xi)$ satisfies condition (9) for every $n \in \mathbb{Z} \setminus \{0\}$ and $r = 0, \dots, L$. Take any infinite path $\pi = \alpha_0 \to \alpha_1 \to \cdots$. Since $\widehat{\varphi}(\xi)$ is holomorphic in the strip Γ (see the proof of Lemma 1) it follows that there exists an integer $k \geq 2$ such that $\widehat{\varphi}(\alpha_{k-1}) \neq 0$. Then equality (10) yields

$$\widehat{\varphi}(2^k \alpha_{k-1}) = \widehat{\varphi}(\alpha_{k-1}) \prod_{l=1}^{k} m_0(2^{k-l}\alpha_{k-1}) = \widehat{\varphi}(\alpha_{k-1}) \prod_{t=0}^{k-1} m_0(\alpha_t).$$

Since $2^k \alpha_{k-1} = 2\pi n$ for some integer $n \neq 0$, it follows that the function $\widehat{\varphi}(\xi)$ has zero of order at least $L+1$ at the point $2^k \alpha_{k-1}$. Consequently the set $\{\alpha_0, \dots, \alpha_{k-1}\}$ contains at least $L+1$ roots of the mask $m_0(\xi)$ (counting with multiplicity). So every infinite path starting at the root of the tree includes at least $L+1$ zeros of m_0. Hence the condition of order L is satisfied.

Conversely, suppose (6) has an L^1-solution $\varphi(x)$ and satisfies the condition of order L for some $L \geq 0$. Take any nonzero integer n. Let $n = 2^k n_1$, where n_1 is an odd integer. If we replace ξ by $2\pi n$ in (11) we obtain

$$\widehat{\varphi}(2\pi n) = \prod_{l=1}^{\infty} m_0(2\pi n 2^{-l}) = m_0(0)^k \prod_{t=0}^{\infty} m_0(\pi n_1 2^{-t}) = \prod_{t=0}^{\infty} m_0(\alpha_t),$$

where α_t is a vertex of the tree such that $\alpha_t \equiv \pi n_1 2^{-t} \pmod{2\pi}$. It follows from the condition of order L that the infinite path $\pi = \alpha_0 \to \alpha_1 \to \cdots$ includes at least $L+1$ zeros of $m_0(\xi)$ (counting with multiplicity). Hence the point $2\pi n$ is a zero of $\widehat{\varphi}(\xi)$ of order at least $L+1$; this completes the proof of Lemma 2. □

Lemmas 1 and 2 imply the following statement: if (6) has an L^1-solution φ such that $\varphi \in W_1^L(\mathbb{R})$ or $s_p(\varphi) > L$, then this equation satisfies the condition of order L. Thus there are $L+1$ minimal cut sets $\mathcal{A}_0, \dots, \mathcal{A}_L$ consisting of roots of equation $m_0(\xi) = 0$. If all these sets are trivial, i.e., $\mathcal{A}_0 = \cdots = \mathcal{A}_L = \{\pi\}$, then the sum rules of order L hold. Otherwise, if (6) is reducible, then at least one of the sets $\mathcal{A}_0, \dots, \mathcal{A}_L$ consists of more than one element. Let $\mathcal{A}_0$ be such a set. Then there is a vertex θ of the tree such that both vertices $\frac{\theta}{2}$ and $\frac{\theta}{2} + \pi$ belong to $\mathcal{A}_0$. It is clear that θ has a form $\theta = \pi(2l+1)2^{-n}$, where $n \geq 0$ is an integer and $l \in \{0, \dots, 2^n - 1\}$. Furthermore, in the case of finitely many terms (equation (1)) we have $n \leq N-2$. Indeed, since the polynomial $m_0(\xi)$ has at most N zeros on the interval $(0, 2\pi)$, we see that $Card\ \mathcal{A}_0 \leq N$ and hence the type of $\mathcal{A}_0$ is at most $N-1$. So the level of the vertex θ is at most $N-2$. This proves item 1 of Theorems 1 and 2. If we prove the next lemma, then the proof of Theorem 2 will be completed.

LEMMA 3. *Let $\varphi(x)$ be an L^1-solution of* (6). *For any $\alpha \in \mathbb{R}$ such that $m_0(\frac{\alpha}{2}) = m_0(\frac{\alpha}{2} + \pi) = 0$, there exists a function $\psi(x)$ such that the following hold:*

(a) $\varphi(x) = \psi(x) - e^{i\alpha}\psi(x-1)$.

(b) *The function $\psi(x)$ satisfies the refinement equation*

$$\psi(x) = \sum_{k \in \mathbb{Z}} \tilde{c}_k \psi(2x - k),$$

where $\tilde{c}_k$ are defined by the equality

$$\sum_{k \in \mathbb{Z}} \tilde{c}_k e^{-i\xi k} = 2\tilde{m}_0(\xi) = 2\frac{m_0(\xi)(1 - e^{i(\alpha - \xi)})}{1 - e^{i(\alpha - 2\xi)}}.$$

(c) *The functions φ and ψ have equal smoothness.*

Proof. Since $\frac{\alpha}{2}$ and $\frac{\alpha}{2}+\pi$ are roots of the mask $m_0(\xi)$, it follows that the function

$$p(\xi) = \sum_{k\in\mathbb{Z}} p_k e^{-ik\xi} = \frac{m_0(\xi)}{1-e^{i(\alpha-2\xi)}}$$

is holomorphic in the strip Γ as well as $m_0(\xi)$ (see the proof of Lemma 1). Therefore the Fourier coefficients $\{p_k\}$ have an exponential decay. Let us define a function ψ by the equality

$$\psi(x) = \sum_{k\in\mathbb{Z}} p_k \varphi(2x-k). \tag{13}$$

It follows easily that ψ is in $L^1(\mathbb{R})$ as well as φ. Further, the Fourier transform of both sides of (13) yields

$$\widehat{\psi}(\xi) = \widehat{\varphi}(\xi/2)p(\xi/2). \tag{14}$$

Hence,

$$\widehat{\varphi}(\xi) = \widehat{\varphi}(\xi/2)m_0(\xi/2) = \widehat{\varphi}(\xi/2)p(\xi/2)(1-e^{i(\alpha-\xi)}) = \widehat{\psi}(\xi)(1-e^{i(\alpha-\xi)}).$$

Thus,

$$\widehat{\varphi}(\xi) = \widehat{\psi}(\xi)(1-e^{i(\alpha-\xi)}). \tag{15}$$

The last equality implies that

$$\varphi(x) = \psi(x) - e^{i\alpha}\psi(x-1). \tag{16}$$

Since the coefficients $\{p_k\}$ have an exponential decay we see that transfer (13) from the function φ to ψ does not decrease the smoothness of function nor does inverse transfer (16). This yields that φ and ψ have an equal smoothness. By the same argument we obtain from formulas (14) and (15) that φ and ψ have an equal Sobolev regularity. Furthermore, using (14) and (15) we have

$$\widehat{\psi}(\xi) = \widehat{\varphi}(\xi/2)p(\xi/2) = \widehat{\psi}(\xi/2)p(\xi/2)(1-e^{i(\alpha-\xi/2)})$$

$$= \frac{\widehat{\psi}(\xi/2)m_0(\xi/2)(1-e^{i(\alpha-\xi/2)})}{1-e^{i(\alpha-\xi)}} = \widehat{\psi}(\xi/2)\tilde{m}_0(\xi/2).$$

Now taking the inverse Fourier transform we obtain

$$\psi(x) = \sum_{k\in\mathbb{Z}} \tilde{c}_k \psi(2x-k).$$

So the function ψ satisfies the refinement equation with the coefficients $\{\tilde{c}_k\}$. Lemma 3 is proved. This completes the proof of Theorem 2. □

Now to conclude the proof of Theorem 1 it remains to establish the additional properties of the function ψ for the case of finitely many terms (see (1)). Since in this case the mask $m_0(\xi)$ is a polynomial of degree N, we see that $p(\xi)$ and $\tilde{m}_0(\xi)$ are polynomials of degree $N-2$ and $N-1$, respectively. Therefore formula (13) can be rewritten for this case as follows:

$$\psi(x) = \sum_{k=0}^{N-2} p_k \varphi(2x-k). \tag{17}$$

We also rewrite the refinement equation for ψ:

$$\psi(x) = \sum_{k=0}^{N-1} \tilde{c}_k \psi(2x-k).$$

Since supp $\varphi = [0, N]$ (see the introduction) it immediately follows from (17) that supp $\psi = [0, N-1]$. Thus in the case of finitely many terms, passing from φ to ψ decreases the order of equation and the support of function. Further, it follows from formula (16) that $\mathcal{V}_\psi \subset \mathcal{V}_\varphi$. On the other hand the inverse transform to (16) is

$$\psi(x) = \sum_{k=0}^{+\infty} e^{i\alpha k} \varphi(x-k).$$

This implies that $\mathcal{V}_\varphi \subset \mathcal{V}_\psi$. Thus $\mathcal{V}_\varphi = \mathcal{V}_\psi$, which completes the proof of Theorem 1. □

Remark 1. Let us note that we actually used the condition $\varphi \in L^1(\mathbb{R})$ only in the proof of formula (11). So it is possible to formulate Theorem 2 in another way, without the assumption $\varphi \in L^1(\mathbb{R})$. Namely, we can define a solution φ of (6) (as in the works [He2] and [CD2]) from equality (11). If the inverse Fourier transform of (11) is a regular reducible function, then Theorem 2 still holds.

Remark 2. The passage from refinement equation with mask $m_0(\xi)$ to the equation with mask $\tilde{m}_0(\xi) = \frac{m_0(\xi)(1-e^{i(\alpha-\xi)})}{1-e^{i(\alpha-2\xi)}}$ is said to be the *transfer to the previous level.* This operation is defined for a mask m_0 if and only if $m_0(\frac{\alpha}{2}) = m_0(\frac{\alpha}{2}+\pi) = 0$. The inverse passage, which is defined for a mask $\tilde{m}_0$ satisfying $\tilde{m}_0(\alpha) = 0$, is said to be correspondingly the *transfer to the next level.* It follows from Lemma 3 that the smoothness of refinable function is stable with respect to both these operations. If α is a vertex of the tree, then the transfer to the next level is, in fact, the passage from α to the next level vertices $\frac{\alpha}{2}$ and $\frac{\alpha}{2}+\pi$. This justifies our terminology. In particular, the decomposition of a reducible function by Theorem 2 is actually a transfer to the previous level.

4. Some conclusions.

Corollary 1 (a necessary condition in terms of roots of the mask). *If refinement equation* (6) *has an L^1-solution that belongs to the space W_1^L , $L \geq 0$, then condition of order L is satisfied.*

Proof. This follows from Lemmas 1 and 2. □

Remark 3. In the case of finitely many terms it is quite easy to verify this necessary condition for a given refinement equation. It is sufficient to look over all the numbers of form $\pi q 2^{-l}$, $q \in \{1, \dots, 2^{l+1}-1\}$, $0 \leq l \leq N-L-1$ (N is the order of the equation), and put those of them that are roots of the polynomial $m_0(\xi)$ onto the tree. Here the multiplicity of any vertex is said to be equal to the multiplicity of the corresponding root of m_0. If this family of vertices has at least $L+1$ common elements (counting with multiplicity) with each infinite path from the root then the condition of order L is satisfied.

Corollary 2 (unimprovability of the necessary condition). *For every family of minimal cut sets $\mathcal{A}_0, \dots, \mathcal{A}_L$, refinement equation* (1) *of order $N = \sum_{k=0}^{L} Card\ \mathcal{A}_k$ having the mask*

$$m_0(\xi) = \frac{1}{2^L} \prod_{k=0}^{L} \prod_{\alpha \in \mathcal{A}_k} (1 - e^{i(\alpha - \xi)})$$

has a compactly supported W_∞^L-solution.

Proof. Applying the transfer to the previous level (Remark 2) we can pass from the mask m_0 to the mask

$$\tilde{m}_0(\xi) = \left(\frac{1 - e^{i(\pi - \xi)}}{2} \right)^{L+1}$$

after several steps. The solution of the corresponding refinement equation is the cardinal B-spline that belongs to W_∞^L (see, for instance, [Sc] or [DL1]). □

Example 1. A necessary condition for the existence of L_1-solution of a refinement equation is the following:

There is a minimal cut set $\mathcal{A}_0$ that consists of roots of the mask.

Remark 4. It is quite surprising that for refinement equations with nonnegative coefficients the necessary condition from Example 1 is sufficient as well. This fact was proved in [W2] (for the case $\mathcal{A}_0 = \{\pi\}$) and in [P] (for the case of arbitrary minimal cut set $\mathcal{A}_0$). Nevertheless, for $L \geq 1$ the condition of order L is not sufficient for the existence of W_1^L-solutions even if all the coefficients of (1) are positive.

Remark 5. It follows from Lemmas 1 and 2 that the condition of order L is also necessary for the property $s_p(\varphi) > L$, where φ is an L^1-solution of corresponding refinement equation.

The next two corollaries concern the Riesz basis property and linear independence of integer translates of refinable functions.

The solution φ of (1) is said to be *l_p-stable* ($1 \leq p \leq \infty$) if $\varphi \in L_p(\mathbb{R})$ and there exist positive constants A_p, B_p such that for any sequence $a = \{a_k\} \in l_p$,

$$A_p \|a\|_p \leq \left\| \sum_{k \in \mathbb{Z}} a_k \varphi(x - k) \right\|_p \leq B_p \|a\|_p; \tag{18}$$

in other words, the integer translates $\{\varphi(x - k),\ k \in \mathbb{Z}\}$ form a Riesz basis of the closure of their linear span in $L_p(\mathbb{R})$. This property is also called the *Riesz basis property* of the function φ.

The l_p-stability of solutions of refinement equations plays an essential role in the study of wavelets (see [Hu], [V], [C], [JW]). This property is also used in problems concerning the convergence of subdivision algorithms [CDM], [DM] and in the study of stability of refinable functions for small perturbations of the coefficients [H]. Several important results on l_p-stability were obtained in [V], [Z], [JW], [He3]. It is a well-known fact that the first sum rule (equality (5)) is a necessary condition for l_p-stability in the case $p = 2$ (see [V]). The following statement generalizes this result to the sum rules of arbitrary order and to arbitrary $p \in [1, \infty]$.

Corollary 3. *Every reducible refinable function is not l_p-stable for any $p \in [1, \infty]$.*

Proof. Theorem 1 implies that for any reducible equation there exists a number $\theta \in \mathbb{R}$ such that $m_0(\frac{\theta}{2}) = m_0(\frac{\theta}{2} + \pi) = 0$. This equality contradicts the Riesz basis property [JW, Theorem 1].

Another important question in the study of refinable functions is linear independence of their integer translates (see [CDM], [C], [JW], [Z]). Theorem 1 yields the following result.

COROLLARY 4. *Integer translates of any reducible function are linearly dependent.*

Proof. We could again use the results of work [JW], but it is simpler to prove this corollary in another way. By Theorem 1 a reducible function $\varphi(x)$ can be decomposed into the sum $\varphi(x) = \psi(x) - e^{i\theta}\psi(x-1)$ for suitable θ and ψ. Whence,

$$\sum_{k\in\mathbb{Z}} e^{ik\theta}\varphi(x-k) = \sum_{k\in\mathbb{Z}} (e^{ik\theta}\psi(x-k) - e^{i(k+1)\theta}\psi(x-k-1)) = 0. \qquad \square$$

As we mentioned in section 2, a lot of previous results on the refinement equations, obtained for the case of complete sum rules, can be transferred by Theorem 1 to the general case without any change. We present one example.

COROLLARY 5. *Any W_1^L-refinable function possesses the following property:*

$$\sum_{k\in\mathbb{Z}} (x-k)^r \varphi(x-k) = \text{const}, \quad r = 0, \dots, L. \tag{19}$$

Remark 6. In the case $L = 0$ this statement was proved independently in [W1] and [CDM] (in that work the case of multivariate continuous refinable functions was considered).

Proof. This statement holds for equations that satisfy the sum rules (see [D]). Since the reduction by Theorem 1 preserves property (19) then this property holds for all refinable functions. $\square$

Corollaries 1 and 5 make it possible to obtain some information about a polynomial space which is spanned by integer translates of a refinable function. The problem is to construct necessary and sufficient conditions on a refinable function φ, which ensure that the space $\mathcal{V}_\varphi$ contains a given polynomial space. This problem has been studied in connection with subdivision algorithms (see [DyL] or [CDM] for many references). The work [CDM] contains a detailed multivariate discussion of this question. In particular it was shown that in the case when the integer translates $\{\varphi(x-k),\ k \in \mathbb{Z}\}$ are linearly independent, $\mathcal{V}_\varphi$ contains the space Π_L if and only if the sum rules of order L are satisfied. (Let us remember that we denote by Π_L the space of polynomials of degree at most L.) In the general case (without the independence assumption) Theorem 9.1 in [CDM] yields several criteria for the inclusion $\Pi_L \subset \mathcal{V}_\varphi$ to hold. Now let us formulate another criterion (for the univariate case) that is simpler to apply. We claim that the condition $\Pi_L \subset \mathcal{V}_\varphi$ is equivalent to the condition of order L (see Definition 1).

THEOREM 3. *For any refinable function $\varphi(x)$ and for any integer $L \geq 0$ the following properties are equivalent:*

(i) *The mask of the corresponding refinement equation* (1) *satisfies the condition of order L;*

(ii) *equality* (19) *holds for every $r = 0, \dots, L$;*

(iii) *for every $r = 0, \dots, L$ the function $\sum_{k\in\mathbb{Z}} k^r \varphi(x-k)$ is a polynomial of degree r. The leading coefficient of this polynomial is equal to* 1*;*

(iv) $\Pi_L \subset \mathcal{V}_\varphi$.

Proof. (i) ⇒ (ii). The proof is the same as that of Corollary 5.

(ii) ⇒ (iii). In the case $L = 0$ equality (19) yields $\sum \varphi(x-k) = \int \varphi(x)dx = 1$. Now property (iii) can be easily established by induction with respect to the parameter L.

(iii) ⇒ (iv). This is obvious.

(iv) ⇒ (i). The property $\Pi_L \subset \mathcal{V}_\varphi$ implies that $\widehat{\varphi}^{(r)}(2\pi n) = 0$ for every $n \in \mathbb{Z} \setminus \{0\}$, $r = 0, \dots, L$ ([CDM, Theorem 9.1]). Now we apply Lemma 2. This concludes the proof of Theorem 3. □

Remark 7. It was shown in [HC] (Proposition 4.13 of that work) that the sum rules of order L are sufficient for a refinable function to possess property (iii). Theorem 3 generalizes this result.

Remark 8. Theorem 3 and Corollary 1 imply in particular that every C^L-refinable function possesses property (iv). This fact was first established in Theorem 8.3 of [CDM].

Remark 9. Note that the fast decay assumption (7) in the statement of Theorem 2 is essential. In general the second theorem of reduction does not hold for refinement equations with infinitely many terms. Namely, there is a refinement equation

$$\varphi(x) = \sum_{k \in \mathbb{Z}} c_k \varphi(2x - k) \tag{20}$$

that is reducible (it has a unique, up to normalization, L_1-solution φ and does not satisfy the first sum rule (5)) and that possesses the following property:

If for some $\theta \in \mathbb{R}$ and $\psi \in L_1$ the decomposition

$$\varphi(x) = \psi(x) - e^{i\theta}\psi(x-1)$$

holds, then the function ψ does not satisfy any refinement equation.

Moreover, the coefficients $\{c_k\}$ of (20) decrease faster than polynomially, i.e., $\sup_{k \in \mathbb{Z}} |c_k k^l| < \infty$ for all $l > 0$. This example was constructed in [P]. Thus the assumption of exponential decay in the statement of Theorem 2 cannot be replaced by one of polynomial decay.

5. The reducibility property on time domain. In this section we restrict ourselves to the case of finitely many terms (refer to (1)). Throughout this paper we use the Fourier transform technique on frequency domain. However, there is another approach in the study of refinement equations with finitely many terms. This approach is based on the use of finite-dimensional linear operators on the time domain. This leads in many aspects to sharper results on refinable functions than the Fourier transform method (see the introduction for references). The question arises whether anything special can be said about reducible refinement equations in terms of the time domain approach. Before we formulate the main result of this section let us recall some notation.

For a given refinement equation (1) consider the two linear operators T_0 and T_1 acting on $\mathbb{C}^N$ and defined by matrices as follows:

$$(T_0)_{ks} = c_{2k-s-1}, \qquad (T_1)_{ks} = c_{2k-s}, \tag{21}$$

where c_l is the coefficient of (1) if $l \in \{0, 1, \dots, N\}$ and $c_l = 0$ otherwise.

Suppose (1) has an L^1-solution $\varphi(x)$. Then this solution is supported in the segment $[0, N]$ (see the introduction). Further, let us define a vector-function v_φ :

$[0,1] \longrightarrow \mathbb{C}^N$ by equality

$$v_\varphi(x) = \Big(\varphi(x), \varphi(x+1), \dots, \varphi(x+N-1)\Big)^T.$$

If φ is a continuous function on $\mathbb{R}$, then for any dyadic $x \in [0,1]$ we have the formula

$$v_\varphi(x) = T_{d_1} \cdots T_{d_q} v_\varphi(0). \tag{22}$$

Here $x = 0.d_1 \cdots d_q$, the index $d_j(x)$ is the jth digit in the binary expansion for x, and $v_\varphi(0)$ is an eigenvector of T_0 with the eigenvalue 1 (see [DL1], [CH]).

Otherwise, if $\varphi(x)$ is not continuous, we can use another formula

$$\int\limits_x^{x+2^{-q}} v_\varphi(x)dx = 2^{-q} T_{d_1} \cdots T_{d_q} \left(\int\limits_0^1 v_\varphi(x)dx \right) \tag{23}$$

instead of equality (22) (see [LW]). Now we are going to establish some special properties of the operators T_0 and T_1 in the case of reducible refinement equations.

First let us introduce some further notation. For a refinement equation denote by $\mathcal{B}$ the set of vertices of the tree that are roots of the mask of this equation. Further, denote by $\tilde{\mathcal{B}}$ the set of the vertices that are blocked by $\mathcal{B}$, i.e.,

$$\alpha \in \tilde{\mathcal{B}} \Longleftrightarrow \text{every infinite path from } \alpha \text{ intersects } \mathcal{B}.$$

For example, the set $\mathcal{B}$ of (8) is $\{\frac{\pi}{2}, \ \frac{3\pi}{2}\}$; consequently $\tilde{\mathcal{B}} = \{\pi\}$. It follows from Lemmas 1 and 2 that the set $\mathcal{B}$ of a reducible equation contains at least one nontrivial (other than $\{\pi\}$) minimal cut set. Hence for any reducible refinement equation the set $\tilde{\mathcal{B}}$ is nonempty.

For any $\beta \in \mathbb{R}$ let us define the vector $u(\beta) = (1, e^{i\beta}, e^{2i\beta}, \dots, e^{i(N-1)\beta})^T \in \mathbb{C}^N$.

As usual we shall denote by $\langle x, y \rangle = \sum_{j=1}^N x_j \overline{y_j}$ the scalar product in $\mathbb{C}^N$, by $Span(M)$ the linear span of a given set M in $\mathbb{C}^N$, by A^* the conjugate operator for a given operator A.

Proposition 1.

(1) Matrices T_0 and T_1 of every reducible refinement equation are degenerate. Moreover, T_0^* and T_1^* have a nontrivial common kernel.

(2) The family of operators $\{T_0^*, T_1^*\}$ is nilpotent on the nontrivial invariant subspace $Span\{u(\alpha), \ \alpha \in \tilde{\mathcal{B}}\}$.

(3) $Span\{v_\varphi(x), \ \ x \in [0,1]\} \perp Span\{u(\alpha), \ \alpha \in \tilde{\mathcal{B}}\}$.

Proof. Item (1) obviously follows from (2). To prove item (2) observe the following property of operators T_0 and T_1, which can be verified by direct calculation:

$$\begin{aligned} T_0^* u(\beta) &= \overline{m_0\Big(\frac{\beta}{2}\Big)} u\Big(\frac{\beta}{2}\Big) + \overline{m_0\Big(\frac{\beta}{2}+\pi\Big)} u\Big(\frac{\beta}{2}+\pi\Big), \\ T_1^* u(\beta) &= e^{-\frac{i\beta}{2}} \overline{m_0\Big(\frac{\beta}{2}\Big)} u\Big(\frac{\beta}{2}\Big) + e^{-i(\frac{\beta}{2}+\pi)} \overline{m_0\Big(\frac{\beta}{2}+\pi\Big)} u\Big(\frac{\beta}{2}+\pi\Big). \end{aligned} \tag{24}$$

These equalities immediately imply that

$$T_{d_1}^* \cdots T_{d_q}^* u(\alpha) = 0 \qquad \text{for every } \alpha \in \tilde{\mathcal{B}} \text{ and } (d_1, \dots, d_q) \in \{0,1\}^q, \tag{25}$$

where q is the type of the set $\mathcal{B}$. So the family $\{T_0^*, T_1^*\}$ is nilpotent on the subspace $Span\{u(\alpha),\ \alpha \in \tilde{\mathcal{B}}\}$. It remains to note that for any reducible equation this subspace is nontrivial, since the set $\tilde{\mathcal{B}}$ is nonempty. This proves item (2).

To prove item (3) assume first that φ is a continuous function on $\mathbb{R}$. It follows from (25) that the equality

$$\Big\langle T_{d_1} \cdots T_{d_q} b,\ u(\alpha) \Big\rangle = 0$$

holds for any $b \in \mathbb{C}^N$. Applying (22) we conclude that equality $\langle v_\varphi(x), u(\alpha) \rangle = 0$ holds for every dyadic x, and consequently, since $v_\varphi(x)$ is continuous, it holds for every $x \in [0, 1]$. So the proof is completed for a continuous φ.

In the other case, when $\varphi \in L^1(\mathbb{R})$ is not continuous, we have to repeat the proof stated above, but we must apply formula (23) instead of (22). □

COROLLARY 6. *If at least one of the matrices T_0, T_1 of* (1) *is nondegenerate then the sum rules of order L are necessary for the existence of a W_1^L-solution.*

Remark 10. Formulas (24) can also be obtained from general properties of the transfer operator (see [E], [CD2] for details). In particular a very similar expression was used in [CD2, formula 7.3].

6. On convergence of the cascade algorithm. In this section we consider equations with finitely many terms (see (1)). In the paper [D] Daubechies introduced the *cascade algorithm* for finding solutions of refinement equations. The one iteration of that algorithm is

$$f_n = T f_{n-1},$$

where

$$Tf(x) = \sum_k c_k f(2x - k). \tag{26}$$

If f_n converges uniformly to a continuous function φ for some initial function f_0, then φ is a continuous solution of corresponding refinement equation (1), and moreover, f_0 possesses the property

$$\sum_k f_0(x - k) \equiv \text{ const} \tag{27}$$

(see [CDM], [Du1]). The cascade algorithm *converges* if f_n converges uniformly to φ for any compactly supported continuous function f_0 satisfying (27). Properties of the cascade algorithm have been studied by many authors in various contexts. The cascade algorithm gives a simple way to approximate refinable functions. In particular this yields applications to the study of wavelets [D], [DL1], [Du1], [Du2]. On the other hand the convergence of the cascade algorithm is equivalent to the convergence of the associated subdivision scheme (see [DM] for many references). It is clear that convergence of the cascade algorithm implies the existence of a continuous solution for the corresponding refinement equation. In general the converse does not hold. (See [DL1] for corresponding examples. The correlation between the existence of continuous solution of a refinement equation and the convergence of the associated subdivision scheme is a rather complicated question. See [CDM], [DM] for general multivariate discussions of this aspect. Some interesting results of the papers [HC] and [CH] actually concern this question, although subdivision schemes were not mentioned explicitly in those works.) It is a well-known fact that the first sum rule (5) is a necessary condition for the cascade algorithm to converge [DGL2], [DyL]. Nevertheless, the sum rules of order $L \geq 1$ are not necessary for that, even if the limit refinable function is in C^L. Corollary 7 of this section yields necessary conditions to the convergence of the cascade algorithm to a smooth function. These conditions are sharp in some sense (Corollary 8). Another problem is to obtain a similar result to that of Theorem 1 for cascade algorithms, i.e., to reduce the general case to the well-studied case of complete sum rules. In general, the convergence property of the cascade algorithm is not stable with respect to the operation of transfer to the next level, unlike the existence and smoothness of refinable functions. This is one example.

Example 2. The cascade algorithm associated to (8) does not converge, since the first sum rule does not hold. Applying the reduction by Theorem 1 we obtain the irreducible equation

$$\varphi(x) = \frac{1}{3}\varphi(2x) + \varphi(2x-1) + \frac{2}{3}\varphi(2x-2). \tag{28}$$

The cascade algorithm associated to this equation converges (this follows directly from Theorem 3.3 of [CDM]). So the convergence property fails with the transfer to the next level from (28) to (8).

Thus the reduction by Theorem 1 can change the convergence property of the cascade algorithm. Before we formulate the main result of this section let us introduce some notation. Consider the space $C_0(\mathbb{R})$ of compactly supported continuous functions on $\mathbb{R}$. A sequence $\{f_n\} \subset C_0$ tends to zero in this space if $f_n \to 0$ uniformly on $\mathbb{R}$ and the supports of the functions f_n are uniformly bounded. Define the subspace $\mathcal{L} \subset C_0$ as follows:

$$\mathcal{L} = \left\{ f \in C_0 \Big| \ \sum_k f(x-k) \equiv 0 \right\}.$$

Note that by the Poisson summation formula the equality $\sum_k f(x-k) \equiv 0$ is equivalent to $\widehat{f}(2\pi n) = 0,\ n \in \mathbb{Z}$. It easily can be proved that the operator T defined by (26) preserves the subspace $\mathcal{L}$ if and only if the first sum rule holds, i.e., $m_0(\pi) = 0$. Moreover, the cascade algorithm converges if and only if the operator T^n restricted to $\mathcal{L}$ tends to zero (in the topology of C_0) as $n \to \infty$.

We shall also need the notion of the transfer to the next (previous) level defined in Remark 2.

LEMMA 4. *Let m_0 be a mask satisfying $m_0(\pi) = 0$. Suppose $\tilde{m}_0$ is obtained from m_0 by a transfer to the previous level; then the cascade algorithm associated to the mask m_0 converges if and only if the algorithm associated to $\tilde{m}_0$ does.*

Proof. Let us define the function $p(\xi)$ and the sequence $\{p_k\}$ as in the proof of Lemma 3:

$$p(\xi) = \sum_{k=0}^{N-2} p_k e^{-ik\xi} = \frac{m_0(\xi)}{1 - e^{i(\alpha - 2\xi)}}.$$

Define also two operators P and S acting in the space C_0:

$$\left(P\psi\right)(x) = \sum_{k=0}^{N-2} p_k \psi(2x-k),$$

$$(S\psi)\,(x) = \psi(x) - e^{i\alpha}\psi(x-1).$$

In the frequency domain these operators have the following form:

$$\widehat{P\psi}(\xi) = \widehat{\psi}(\xi/2)p(\xi/2); \quad \widehat{S\psi}(\xi) = \widehat{\psi}(\xi)(1-e^{i(\alpha-\xi)}). \tag{29}$$

Now observe the following relations:

$$PS = \tilde{T}, \quad SP = T, \tag{30}$$

where T and $\tilde{T}$ are operators of the cascade algorithms associated to m_0 and $\tilde{m}_0$, respectively. To prove this we apply (29) and get

$$\widehat{PS\psi}(\xi) = p(\xi/2)(1-e^{i(\alpha-\xi/2)})\widehat{\psi}(\xi/2) = \tilde{m}_0(\xi/2)\widehat{\psi}(\xi/2) = \widehat{\tilde{T}\psi}(\xi).$$

The equality $SP = T$ can be established in the same way. [It is clear that both P and S are continuous operators on $C_0(\mathbb{R})$.] Furthermore, both P and S preserve the subspace $\mathcal{L}$. To see this take any $\psi \in \mathcal{L}$ and apply formulas (29)

$$\widehat{P\psi}(2\pi n) = \widehat{\psi}(\pi n)p(\pi n) = 0.$$

Let us clarify the last equality. If n is even, then $\widehat{\psi}(\pi n) = 0$, since $\psi \in \mathcal{L}$. If n is odd then by the assumption $p(\pi n) = 0$. For the operator S we have

$$\widehat{S\psi}(2\pi n) = \widehat{\psi}(2\pi n)(1-e^{i(\alpha-2\pi n)}) = 0.$$

It follows from (30) that $T^k = S\tilde{T}^{k-1}P$ for every $k \geq 1$. Therefore, since P and S are continuous and preserve the subspace $\mathcal{L}$, the convergence $\tilde{T}^{k-1} \to 0$ on $\mathcal{L}$ implies the convergence $T^k \to 0$ on $\mathcal{L}$. Conversely, from the equality $\tilde{T}^k = PT^{k-1}S$ it follows that the convergence $T^{k-1} \to 0$ it implies that $\tilde{T}^k \to 0$. The lemma is proved. ∎

As a corollary we obtain the following proposition that reduces the problem of convergence of the cascade algorithms to the case of complete sum rules.

PROPOSITION 2. Let us have a reducible refinement equation satisfying the first sum rule. Suppose we pass to an irreducible equation by Theorem 1 after several steps; then the cascade algorithm associated to the first equation converges if and only if the algorithm associated to the second equation does.

Another corollary yields a necessary condition for the cascade algorithm to converge to a function from $W_1^L(\mathbb{R})$.

COROLLARY 7. *If the cascade algorithm associated to a mask m_0 converges to a compactly supported function from W_1^L $(L \geq 1)$ then the condition of order L is satisfied and $m_0(\pi) = 0$, i.e., there are $L+1$ minimal cut sets of the tree consisting of roots of the mask (counting with multiplicity), and one of them is trivial (coincides with $\{\pi\}$).*

This condition is sharp in the sense that every possible case is realized. The proof of the following corollary is the same as that of Corollary 2.

COROLLARY 8 (unimprovability of the necessary condition). *For every family of minimal cut sets $\mathcal{A}_0, \dots, \mathcal{A}_L$, where $L \geq 1$ and $\mathcal{A}_0 = \{\pi\}$, the cascade algorithm associated to the mask*

$$m_0(\xi) = \frac{1}{2^L}\prod_{k=0}^{L}\prod_{\alpha\in\mathcal{A}_k}(1-e^{i(\alpha-\xi)})$$

converges to a compactly supported function from W_∞^L.

Acknowledgments. I would like to sincerely thank Professors I. Daubechies, S. Konyagin, and J. Lagarias for helpful discussions and Professor Yu. Farkov for useful remarks.

This paper was written when I was visiting the Institute for Advanced Study. I am very grateful to the Institute for its warm hospitality.

REFERENCES

[CDM] D. Cavaretta, W. Dahmen, and C. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[C] C.K. Chui, *An introduction to wavelets,* Wavelet Anal. Appl., 1, Academic. Press., New York, 1992, pp. 1–267.

[CD1] A. Cohen and I. Daubechies, *A stability criterion for the orthogonal wavelet bases and their related subband coding scheme*, Duke Math. J., 68 (1992), pp. 313–335.

[CD2] A. Cohen and I. Daubechies, *A new technique to estimate the regularity of refinable functions*, Rev. Mat. Iberoamericana, 12 (1996), pp. 527–591.

[CH] D. Colella and C. Heil, *Characterization of scaling functions: Continuous solutions,* SIAM J. Matrix Anal. Appl., 15 (1994), pp. 496–518.

[DM] W. Dahmen and C.A. Micchelli, *Biorthogonal wavelets expansion,* Constr. Approx., 13 (1997), pp. 293–328.

[D] I. Daubechies, *Orthonormal bases of wavelets with compact support*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[DL1] I. Daubechies and J.C. Lagarias, *Two-scale difference equations.* I. *Existence and global regularity of solutions,* SIAM. J. Math. Anal., 22 (1991), pp. 1388–1410.

[DL2] I. Daubechies and J.C. Lagarias, *Two-scale difference equations.* II. *Local regularity, infinite products of matrices and fractals,* SIAM. J. Math. Anal., 23 (1992), pp. 1031–1079.

[DDL] G.A. Derfel, N. Dyn, and D. Levin, *Generalized refinement equations and subdivision processes*, J. Approx. Theory, 80 (1995), pp. 272–297.

[DD] G. Deslauriers and S. Dubus, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.

[Du1] S. Durand, *Convergence of the cascade algorithms introduced by I. Daubechies*, Numer. Algorithms, 4 (1993), pp. 307–322.

[Du2] S. Durand, *Etude de la vitesse de convergence de l'algorithme en cascade dans la construction des ondeletters d'Ingrid Daubechies*, Rev. Mat. Iberoamericana, 12 (1996), pp. 277–297.

[DGL1] N. Dyn, J.A. Gregory, and D. Levin, *A four-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Design, 4 (1987), pp. 257–268.

[DGL2] N. Dyn, J.A. Gregory, and D. Levin, *Analysis of linear binary subdivision schemes for curve design*, Constr. Approx., 7 (1991), pp. 127–147.

[DyL] N. Dyn and D. Levin, *Interpolatory subdivision schemes for the generation of curves and surfaces*, in Multivariate Approximation and Interpolation, Duisburg, 1989, Birkhäuser, Basel, 1990, pp. 91–106.

[E] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal., 23 (1992), pp. 1015–1030.

[H] C. Heil, *Some stability properties of wavelets and scaling functions*, NATO Adv. Sci. Inst. Ser. C. Math. Phys. Sci., 442, Kluwer Academic Publishing, Dordrecht, The Netherlands, 1994, pp. 19–38.

[HC] C. Heil and D. Collela, *Dilation equations and the smoothness of compactly supported wavelets*, in Wavelets: Mathematics and Applications, J. Benedetto and M. Frazier, eds., CRC Press, Boca Raton, FL, 1993, pp. 161–200.

[He1] L. Herve, *Régularité et conditions de bases de Riesz por les fonctions d'échelle*, C. R. Acad. Sci. Paris Ser. I Math., 335 (1992), pp. 1029–1032.

[He2] L. Herve, *Construction et régularité des fonctions d'échelle,* SIAM J. Math. Anal, 26 (1995), pp. 1361–1385.

[He3] L.Herve, *Etude d'opérateurs quasy-compacts positifs. Applications aux opérateurs de transfert*, Ann. Inst. H. Poincaré Probab. Statist., 30 (1994), pp. 437–466.

[Hu] Y. Huang, *A nonlinear operator related to scaling functions and wavelets*, SIAM J. Math. Anal., 27 (1996), pp. 1770–1790.

[JW] R.Q. Jia and J. Wang, *Stability and linear independence associated with wavelet decom-*

position, Proc. Amer. Math. Soc., 117 (1993), pp. 1115–1124.

[LMW] K.-S. Lau, M.-F. Ma, and J. Wang, *On some sharp regularity estimations of L^2-scaling functions,* SIAM. J. Math. Anal., 27 (1996), pp. 835–864.

[LW] K.S. Lau and J. Wang, *Characterization of L_p-solutions for two-scale dilation equations,* SIAM. J. Math. Anal., 26 (1995), pp. 1018–1046.

[M] C.A. Micchelli, *Subdivision algorithms for curves and surfaces*, in Proceedings, Extensions of B-Spline Curve Algorithms to Surfaces, Course 5, SIGGRAPH, Dallas, TX, 1986.

[MP] C.A. Micchelli and H. Prautzsch, *Uniform refinement of curves*, Linear Algebra Appl., 114/115 (1989), pp. 841–870.

[P] V. Protasov, *Refinement equations with nonnegative coefficients*, J. Fourier Anal. Appl., to appear.

[Sc] L.L. Schumaker, *Spline functions: Basic theory,* John Wiley, New York, 1981.

[V] 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.

[W1] Y. Wang, *Two-scale dilation equations and the cascade algorithm*, Random Comput. Dynam., 3 (1995), pp. 289–307.

[W2] Y. Wang, *Two-scale dilation equations and the mean spectral radius*, Random Comput. Dynam., 4 (1996), pp. 49–72.

[Z] D.-X. Zhou, *Stability of refinable functions, multiresolution analysis, and Haar bases*, SIAM J. Math. Anal., 27 (1996), pp. 891–904.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1351–1373

THE MECHANISM OF THE POLARIZATIONAL MODE INSTABILITY IN BIREFRINGENT FIBER OPTICS*

YI A. LI† AND KEITH PROMISLOW‡

Abstract. We show that the soliton solutions of the integrable Manakov equation exhibit an instability under arbitrarily small Hamiltonian perturbations. The instability arises from eigenvalues embedded in the essential spectrum of the associated linearized operators; these eigenvalues are dislodged by smooth perturbations. Specifically we consider perturbations which arise in fiber optics as a result of birefringence, including the so-called four-wave mixing term. Employing the Evans function and a Dirichlet expansion on the stable manifold of the linearized system, we obtain rigorous perturbation results and compute the stability diagram of the fast wave for all physical values of the birefringent parameters, using a novel numerical scheme derived from the Dirichlet expansion.

Key words. traveling waves, Evans function, polarization mode instability, Dirichlet expansion

AMS subject classifications. 34A47, 34C35, 34D30, 35P20, 35Q55, 78A60

PII. S0036141099349966

1. Introduction.

Soliton pulses form the backbone of high speed fiber-optic telecommunication systems and hold great potential for all optical switching devices. Attempts to further increase bit rates exploit the robust elastic collision properties of solitons, in particular, novel soliton packing schemes propose to subdivide information streams into different wavelengths—wavelength division multiplexing (WDM), or into orthogonal polarizations—polarization division multiplexing (PDM) [9]. However, a proper modeling of these schemes, especially to address stability properties, requires careful attention to the various perturbations which are present in optical systems.

To model the interaction of the orthogonal polarizations of a pulse, the widely studied nonlinear Schrödinger equation (NLS) is extended to the integrable Manakov system. However, fiber-optic systems exhibit birefringence, differing phase and group velocities for different polarizations, as well as nonlinear interactions between polarizations dependent upon amplitude—cross-phase modulation (XPM) and upon complex phase—four-wave mixing (FWM). The nonlinear terms break the integrability of the Manakov equation and numerical simulations including these nonlinearities have demonstrated pulse splitting and inelastic collisions, [32, 38]. We consider the case of weak birefringence, neglecting group velocity birefringence and higher order dispersions.

Many applications of single-mode fibers require transmission of a stable state of polarization. In PDM, to reduce tail-tail interactions and increase soliton packing, solitons are sent in sequences with polarizations alternating between the fast and slow polarization axes. While the slow wave enjoys all the stability properties of the NLS soliton, in the nonintegrable case it was discovered numerically [6] that the fast wave can experience what is termed the polarization mode instability, which leads to loss of linear polarization. It has been common to neglect the FWM, arguing that the polarization axes change rapidly or even stochastically with propagation distance.

*Received by the editors January 18, 1999; accepted for publication (in revised form) September 23, 1999; published electronically May 19, 2000.

http://www.siam.org/journals/sima/31-6/34996.html

†CNLS, Los Alamos National Laboratory, Los Alamos, NM 87545 (yili@cs.stevens-tech.edu).

‡Department of Mathematics and Statistics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (kpromisl@sfu.ca). The second author was supported by NSERC grant 611255.

However, for soliton collisions and applications where polarization holding is desired, the polarization axes can be assumed constant with distance. Moreover, we show rigorously that the FWM produces sensitive, leading-order effects on the stability of the fast wave.

The linearization of the perturbed equations about the fast wave yields a two-parameter family of non-self-adjoint operators with a rich structure. In the integrable case there are eigenvalues embedded in the essential spectrum, and the stability of the fast wave depends upon the fate of these embedded eigenvalues under the continuous perturbations of XPM and FWM. In a beautiful construction, Grillakis [12] showed that arbitrarily small, relatively compact perturbations could eject embedded eigenvalues and produce instability. However, the construction is in a sense unnatural, involving delta functions in the spectral projection, and it has been open to speculation whether such eigenvalues could be ejected by continuous terms. We show that this is exactly the case—arbitrarily small perturbations do eject the embedded eigenvalues, whose distance to the essential spectrum initially grows quadratically in the coefficient of FWM. This is the mechanism of the polarizational mode instability, and it is in this sense that PDM is structurally unstable. We also find edge bifurcations, a special case in which the branch points of the essential spectrum eject eigenvalues.

Our analysis combines two analytical tools in a novel way—the Evans function for the linearization and a Dirichlet expansion on the stable manifold of the associated first order eigenvalue problem. The Evans function is an analytic function whose zeros coincide with the eigenvalues of the operator up to algebraic multiplicity. It has recently been the focus of considerable attention, having proven its effectiveness as an analytical tool for eigenvalue problems; see [16, 18] and the references therein. The works just cited, and particularly [16], develop techniques to calculate the first nonzero derivative of the eigenvalues with respect to the bifurcation parameters. In the analysis at hand we require the second nonzero derivative, and this is the first such computation known to the authors. Indeed, the detailed description of the motion of the eigenvalues, provided in that which follows, is testimony to efficacy of the Evans function. In previous work [22], we investigated the issues of structural stability of these waves, but the functional analytic techniques employed failed to shed light upon the fate of the embedded eigenvalues.

The Evans function and Dirichlet expansion are a particularly effective combination in the near-integrable regime; it also admits analysis when the perturbations are not small. Indeed, even far from the integrable case, the Dirichlet expansion gives a constructive formulation of the Evans function, rendering a transparent analytic extension into the essential spectrum. We exploit this construction to develop a hybrid numerical scheme which efficiently yields an accurate computation of the Evans function, making possible a detailed resolution of the eigenvalue problem. Indeed, we present a sequence of bifurcations common to several families of operators arising as linearizations about traveling waves of integrable systems under Hamiltonian perturbations [2, 26, 31]. Consider a linear operator whose spectrum is symmetric about the real and imaginary axes, and whose essential spectrum, of positive Krein sign (see [12]) takes the form $\{\pm i\mu | \mu > \mu_0\}$ with branch points at $\pm i\mu_0$, leaving uncovered a neck $(-i\mu_0, i\mu_0)$ of the imaginary axis containing the origin. Negative Krein sign eigenvalues embedded in the essential spectrum may bifurcate out under the influence of the perturbative terms. Symmetric groups of four eigenvalues, $\{\lambda, \lambda^*, -\lambda, -\lambda^*\}$, are ejected from the essential spectrum and pass through the complex plane with increasing values of the bifurcation parameter. Depending upon the length of the exposed neck, the eigenvalues may recombine in pairs either upon the real axis or upon the exposed neck of the imaginary axis, or, in a critical case, all four may arrive at the origin simultaneously. We call this phenomena a *neck bifurcation*; see Figures 1 and 2 and the discussion in section 5 for a detailed description. For the underlying equation, this bifurcation may manifest itself as an oscillatory instability under small perturbation, stability under moderate perturbations, and with yet larger perturbations leading to nonoscillatory instability.

This paper is organized as follows. In section 2 we introduce the coupled nonlinear Schrödinger systems (CNLS), the traveling waves, and their associated linearized operators, recalling the previously obtained results. In section 3 the eigenvalue problem is set up and the Dirichlet expansion solutions are constructed. The essence of our analysis lies in section 4, where we introduce the Evans function and obtain the main results, Theorems 4.3 and 4.5, which describe the asymptotic motion of the embedded eigenvalues under small perturbations. In section 5 we address the far from integrable case and provide new, rigorous-numerical results on the stability in this case; in particular, we trace the parameter dependence of the unstable eigenvalues, find the structurally stable regimes, and introduce the *neck bifurcation*.

Remark. Recently, Kapitula and Rubin [17] independently and simultaneously developed an expansion of the Evans function at its branch points similar to the expansion used herein to prove Proposition 4.6. They apply the expansion to detect edge bifurcations undergone by dark soliton solutions of NLS and complex Ginzburg–Landau under a wide class of pertubations.

2. The CNLS equations and notation.

Pulse propagation in linearly birefringent, lossless fibers is described by the CNLS equations [9]

$$\begin{aligned} i(u_t+\delta u_x)+\tfrac{1}{2}u_{xx}-\kappa u+(|u|^2+A|v|^2)u+Bv^2u^* &= 0,\\ i(v_t-\delta v_x)+\tfrac{1}{2}v_{xx}+\kappa v+(A|u|^2+|v|^2)v+Bu^2v^* &= 0, \end{aligned} \tag{2.1}$$

where the coefficient $\kappa>0$ is the phase velocity differential between the two polarizations, the XPM and FWM coefficients, respectively A and B, are positive and satisfy $A+B=1$. The complex valued solution $(u,v)^T$ represents the two orthogonal polarizations of the electromagnetic field, * denotes complex conjugation, and a superscript ^{T} denotes a transpose. Following the optics convention, t connotes scaled distance of propagation and x connotes scaled time. In that which follows, we eliminate the XPM coefficient by the substitution $A=1-B$ and employ the FWM coefficient, B, as the bifurcation parameter.

In what is termed the weak birefringence limit we neglect the group velocity birefringence δ and assume that κ is relatively small. In this form the CNLS equations possess a Hamiltonian $\mathcal{H}=\mathcal{H}_0+\mathcal{H}_1$ where $\mathcal{H}_0$ is the Hamiltonian of the integrable Manakov equation,

$$\mathcal{H}_0=\frac{1}{4}\int_{-\infty}^{\infty}\left[\left(|u_x|^2+|v_x|^2\right)-\left(|u|^2+|v|^2\right)^2+2\kappa\left(|u|^2-|v|^2\right)\right]dx,$$

and $\mathcal{H}_1$ is the birefringent perturbation,

$$\mathcal{H}_1=-\frac{B}{4}\int_{-\infty}^{\infty}\left(vu^*-uv^*\right)^2dx.$$

Large classes of solitary wave solutions of (2.1) have been found explicitly; they are the bound n-soliton solutions of the integrable Manakov equation with n-identical speeds c. We focus attention on the one-soliton. For the Manakov equation the general one-soliton is

$$\begin{pmatrix} u \\ v \end{pmatrix}(x,t) = e^{ic(x-ct)}\eta \begin{pmatrix} e^{i\omega_1 t}\cos\alpha \\ e^{i\omega_2 t}\sin\alpha \end{pmatrix} \phi\left(\eta(x-ct)\right) \tag{2.2}$$

with $\phi(x) = \operatorname{sech} x$, $\eta \geq 0$, and ω_1 and ω_2: related through η, κ, and speed c:

$$\begin{aligned} \omega_1 &= \frac{\eta^2+c^2}{2} - \kappa, \\ \omega_2 &= \frac{\eta^2+c^2}{2} + \kappa. \end{aligned}$$

Due to the Galilean invariance, it is sufficient to consider standing waves with speed $c = 0$. When $B \neq 0$ the CNLS equation does not support the traveling waves with differing phase velocity, and the persistent solutions are $(u,v)^T = e^{i\omega_j t}\Phi_j(x)$, where

$$\begin{aligned} \Phi_1 &= \eta \begin{pmatrix} 1 \\ 0 \end{pmatrix} \phi(\eta x), \\ \Phi_2 &= \eta \begin{pmatrix} 0 \\ 1 \end{pmatrix} \phi(\eta x). \end{aligned}$$

We call these the "up" and "down" waves, respectively, in analogy to the "up" and "down" states of a pendulum. It is well known that the down wave, or slow wave as it is called in the optics community, is the ground state of the Hamiltonian system and is orbitally stable; see [13]. The up wave, or fast wave, is not a ground state and may be linearly unstable. Our goal is to lay transparent the mechanism which leads to the instability of the up wave as B changes from zero.

Addressing the linear stability of the up wave, we rescale as $\tilde{x} = \eta x$ and drop the $\tilde{}$ and the subscript from the phase velocity $\omega = \omega_1$. We consider solutions W to (2.1) which are perturbations of the up wave of the form

$$W = e^{i\omega t}\left(\Phi_1 + e^{\lambda t}U\right),$$

where $U = (u,v)^T$ is small. The rescaling introduces a dimensionless parameter $\rho = \sqrt{\frac{\omega-\kappa}{\omega+\kappa}}$, which takes values in $(0,1)$; ρ expresses the relative strength of the phase velocity differential κ to the phase velocity ω. Keeping terms first order in U, we obtain two uncoupled eigenvalue problems, one each for the two components $u = u_1 + iu_2$ and $v = v_1 + iv_2$, of U.

$$\begin{aligned} J\mathcal{L}_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \lambda \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, && \text{(a)} \\ J\mathcal{L}_2 \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} &= \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, && \text{(b)} \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} \mathcal{L}_1 &= \begin{pmatrix} -\frac{1}{2}\frac{d^2}{dx^2} + \frac{1}{2} - 3\phi^2 & 0 \\ 0 & -\frac{1}{2}\frac{d^2}{dx^2} + \frac{1}{2} - \phi^2 \end{pmatrix}, \\ \mathcal{L}_2 &= \begin{pmatrix} -\frac{1}{2}\frac{d^2}{dx^2} + \frac{\rho^2}{2} - \phi^2 & 0 \\ 0 & -\frac{1}{2}\frac{d^2}{dx^2} + \frac{\rho^2}{2} - (1-2B)\phi^2 \end{pmatrix}, \end{aligned}$$

and $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. The non-self-adjoint operators $L_1 = J\mathcal{L}_1$ and $L_2 = J\mathcal{L}_2$ are disjoint parts of the linearized operator

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix};$$

the spectrum of L is the union of the spectra of L_1 and L_2. The operator L_1 is well known as it arises in the linearization of the cubic nonlinear Schrödinger equation,

$$iu_t + \frac{1}{2}u_{xx} - \kappa u + |u|^2 u = 0,$$

about the solitary wave solution $u = \eta e^{i\omega t}\operatorname{sech}\eta x$. It has only one discrete eigenvalue $\lambda = 0$, with algebraic multiplicity four and a continuous spectrum composed of the line segment $\{\Re z = 0, |\Im z| \geq \frac{1}{2}\}$. To determine if the spectrum of L is purely imaginary it suffices to study the two-parameter family of operators $L_2(\rho, B)$; we consider values $(\rho, B) \in (0,1) \times \mathbf{R}$.

The operator L_2 was studied by the authors in [22]; in particular, it was shown that the spectrum of L_2 is symmetric about both the real and the imaginary axes and the essential spectrum consists of the set $\{\Re z = 0, |\Im z| \geq \frac{\rho^2}{2}\}$. For $B \geq 0$, the point spectrum of L_2 consists of at most four eigenvalues, also symmetric about the real and the imaginary axes. For the integrable case, $B = 0$, L_2 has eigenvalues $\lambda_\pm = \pm i\frac{1-\rho^2}{2}$ with associated eigenvectors $(\phi, \mp i\phi)^T$. For $\rho \leq \frac{1}{\sqrt{2}}$, these eigenvalues are embedded in the essential spectrum. Moreover (see Theorem 6 of [22]) L_2 has no kernel except for critical values $B = B_c(\rho) = \frac{1}{4}(2+\rho)(1-\rho)$ for which zero is an eigenvalue of L_2 of multiplicity two or four. For fixed ρ small enough, and $B = B_c$, the multiplicity is two and the eigenvalues arrive at zero from the real axis as B increases to $B_c(\rho)$. For ρ close enough to 1, the multiplicity is again two but the eigenvalues arrive at zero along the imaginary axis as B increases to B_c, and for at least one value of ρ, the multiplicity is four and the eigenvalues $\{\lambda, \lambda^*, -\lambda, -\lambda^*\}$ arrive symmetrically, one from each quadrant of the complex plane.

In the next two sections we extend these results significantly, localizing the point spectrum of L_2 by computing leading order asymptotics of its eigenvalues for B small, in particular at the embedded eigenvalue and the branch points. The first step is the development of a Dirichlet expansion, given by Theorem 3.2, which constructs solutions of the linearized eigenvalue problem (2b) satisfying one-sided boundary conditions.

3. The reduced system.

We begin by writing the eigenvalue problem for L_2 as a system of first order differential equations

$$Y' = A(\lambda, x)Y, \tag{3.1}$$

where $'$ denotes differentiation with respect to x, $Y = (y_1, y_2, y_3, y_4)^T = (v_1, v_2, v_{1x}, v_{2x})^T$, and

$$A(\lambda, x) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \rho^2 - 2\phi^2 & 2\lambda & 0 & 0 \\ -2\lambda & \rho^2 - 2(1-2B)\phi^2 & 0 & 0 \end{pmatrix}.$$

Since ϕ decays to 0 as $x \to \pm\infty$, the matrix A has asymptotic form A_0 given by

$$A_0(\lambda) = \lim_{|x|\to\infty} A(\lambda, x) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \rho^2 & 2\lambda & 0 & 0 \\ -2\lambda & \rho^2 & 0 & 0 \end{pmatrix},$$

which has eigenvalues $\{\mu_1, -\mu_1, \mu_2, -\mu_2\}$ given by

$$\begin{array}{rcl} \mu_1(\lambda) & = & \sqrt{\rho^2 + 2i\lambda}, \\ \mu_2(\lambda) & = & \sqrt{\rho^2 - 2i\lambda}. \end{array} \tag{3.2}$$

The branch cuts for μ_1 and μ_2 are taken as $-\pi < \arg(\lambda - i\frac{\rho^2}{2}) < \pi$ and $-\pi < \arg(\lambda + i\frac{\rho^2}{2}) < \pi$, respectively. Consequently, for $\lambda \in \Omega = \{z | \Re z > 0 \text{ or } |\Im z| < \frac{\rho^2}{2}\}$ the eigenvalues μ_1 and μ_2 have positive real part. To each of the eigenvalues $\mu_1, \mu_2, \mu_3 = -\mu_1$, and $\mu_4 = -\mu_2$ we associate, respectively, the vector $\vec{\eta}_1 = (1, i, -\mu_1, -i\mu_1)^T$, $\vec{\eta}_2 = (1, -i, -\mu_2, i\mu_2)^T, \vec{\eta}_3 = (1, i, \mu_1, i\mu_1)^T$, $\vec{\eta}_4 = (1, -i, \mu_2, -i\mu_2)^T$ and solutions Y_j of (3.1) for $j = 1, 2, 3, 4$ which satisfy the asymptotic boundary conditions

$$\lim_{x\to x_j} e^{\mu_j x} Y_j(x) = \vec{\eta}_j, \tag{3.3}$$

where $x_j = \infty$ for $j = 1, 2$ and $x_j = -\infty$ for $j = 3, 4$. In particular, for $\lambda \in \Omega$, Y_j decays exponentially as $x \to \infty$ for $j = 1, 2$ and Y_j decays exponentially as $x \to -\infty$ for $j = 3, 4$. We also exploit a symmetry of the family of matrices $A(\lambda)$ which relates the Y_j to each other in a simple way. Define

$$S_1 = \begin{pmatrix} I_{2\times2} & 0 \\ 0 & -I_{2\times2} \end{pmatrix} \text{ and } S_2 = \begin{pmatrix} \sigma_3 & 0 \\ 0 & \sigma_3 \end{pmatrix},$$

where $I_{2\times2}$ is the 2×2 identity matrix and

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

is the third Pauli spin matrix. The following lemma then holds.

Lemma 3.1. *If for each λ and B, $Y(x, \lambda, B)$ solves* (3.1)*, then $S_1 Y(-x, \lambda, B)$ $S_2 Y(x, -\lambda, B)$, and $S_1 S_2 Y(-x, -\lambda, B)$ are also solutions of* (3.1) *for the same values of λ and B. In particular*

$$\begin{array}{rcl} S_1 Y_1(-x, \lambda, B) & = & Y_3(x, \lambda, B), \\ S_1 Y_2(-x, \lambda, B) & = & Y_4(x, \lambda, B), \end{array}$$

and

$$\begin{array}{rcl} S_2 Y_1(x, -\lambda, B) & = & \left\{ \begin{array}{ll} Y_2(x, \lambda, B) & \text{if } \Im\lambda > -\frac{\rho^2}{2}, \\ Y_4(x, \lambda, B) & \text{if } \Im\lambda < -\frac{\rho^2}{2}, \end{array} \right. \\ S_1 S_2 Y_1(-x, -\lambda, B) & = & \left\{ \begin{array}{ll} Y_4(x, \lambda, B) & \text{if } \Im\lambda > -\frac{\rho^2}{2}, \\ Y_2(x, \lambda, B) & \text{if } \Im\lambda < -\frac{\rho^2}{2}. \end{array} \right. \end{array}$$

Proof. These relations follow from the fact that the matrix $A(\lambda)$ verifies the equalities

$$S_1^{-1} A(\lambda) S_1 = -A(\lambda) \text{ and } S_2^{-1} A(\lambda) S_2 = A(-\lambda)$$

from the fact that $S_1\vec{\eta}_j = \vec{\eta}_{j+2}$ for $j = 1, 2$, $S_2\vec{\eta}_j(\lambda) = \vec{\eta}_{j+1}(-\lambda)$ for $j = 1, 3$, and from the relation

$$\mu_1(-\lambda) = \begin{cases} \mu_2(\lambda) & \text{if } \Im\lambda > -\frac{\rho^2}{2}, \\ -\mu_2(\lambda) & \text{if } \Im\lambda < -\frac{\rho^2}{2}. \end{cases} \qquad \square$$

In the following theorem we construct solutions of (3.1) with the prescribed asymptotic behavior.

THEOREM 3.2. *Let λ and B be given. There exists a solution $Y_1(x; \lambda, B)$ of* (3.1) *satisfying*

$$\lim_{x\to\infty} e^{\mu_1 x} Y_1 = \vec{\eta}_1.$$

Moreover $Y_1 = (v_1, v_2, v_{1x}, v_{2x})$, where $V = (v_1, v_2)^T$ is given by the Dirichlet expansion

$$V = e^{-\mu_1 x} \sum_{k=0}^{\infty} \vec{\xi}_k e^{-2kx}, \tag{3.4}$$

uniformly convergent for $x \geq x_0$ for any $x_0 > 0$. The vectors $\vec{\xi}_k$ are defined recursively below. In particular, when $B = 0$ we have the explicit solution

$$V = \frac{\mu_1 + \tanh x}{1 + \mu_1} e^{-\mu_1 x} \begin{pmatrix} 1 \\ i \end{pmatrix}. \tag{3.5}$$

Proof. We rewrite the Dirichlet expansion (3.4) in terms of $z = e^{-x}$, expand $\phi^2(x) = \text{sech}^2 x$ as

$$\phi^2(z) = -4 \sum_{k=1}^{\infty} (-1)^k k z^{2k},$$

and substitute the expansion into (2.3b), resulting in the system of equations

$$\sum_{n=0}^{\infty} \mathcal{A}_n \vec{\xi}_n z^{2n} = -4 \begin{pmatrix} 1 & 0 \\ 0 & (1-2B) \end{pmatrix} \sum_{n=0}^{\infty} \sum_{k=1}^{n} (-1)^k k \vec{\xi}_{n-k} z^{2n}, \tag{3.6}$$

where the matrices $\mathcal{A}_n$ are given by

$$\mathcal{A}_n = \begin{pmatrix} \frac{1}{2}(\rho^2 - (\mu_1 + 2n)^2) & \lambda \\ -\lambda & \frac{1}{2}(\rho^2 - (\mu_1 + 2n)^2) \end{pmatrix}.$$

The matrix $\mathcal{A}_0$ is singular, while

$$\det\, \mathcal{A}_n = n(\mu_1 + n)(2n + \mu_1 + \mu_2)(2n + \mu_1 - \mu_2),$$

and a routine calculation shows that for the branch cuts of μ_1 and μ_2 chosen as above, $\det\mathcal{A}_n$ is nonzero for all complex λ and all $n \geq 1$.

Equating the coefficients of z^0 on both sides of relation (3.6) above, there obtains the equation

$$\mathcal{A}_0 \vec{\xi}_0 = \vec{0},$$

for which we choose the solution

$$\vec{\xi}_0 = \begin{pmatrix} 1 \\ i \end{pmatrix}, \tag{3.7}$$

while equating coefficients of z^n for $n = 1, 2, \ldots$ leads to the following recursive formula for $\vec{\xi}_n$:

$$\vec{\xi}_n = -4\mathcal{A}_n^{-1} \begin{pmatrix} 1 & 0 \\ 0 & (1-2B) \end{pmatrix} \sum_{k=1}^{n} (-1)^k k \vec{\xi}_{n-k}. \tag{3.8}$$

For fixed λ and large n, the matrix $\mathcal{A}_n$ is diagonally dominant with diagonal elements $O(n^2)$ so that

$$\|\mathcal{A}_n^{-1}\| \leq \frac{C}{n^2},$$

for all $n > 0$ and some constant $C = C(\lambda)$. We assume that $|\vec{\xi}_k| \leq Mk^p$ for some $p > 0$ and all positive integers $k < n$ and show that $|\vec{\xi}_n| \leq Mn^p$ if p is large enough, independent of n. This establishes the polynomial growth of the $\vec{\xi}_n$ in n and the uniform convergence of the Dirichlet expansion for $x \geq x_0$ for any $x_0 > 0$. Apply the triangle inequality, the bound on $\|\mathcal{A}_n^{-1}\|$, and the inductive hypothesis to (3.8) to obtain

$$\begin{aligned} \frac{|\vec{\xi}_n|}{n^p} &\leq \frac{4CM}{n^{2+p}} \sum_{k=1}^{n} k(n-k)^p, \\ &\leq 4CM \sum_{k=1}^{n} \tfrac{k}{n} \left(\tfrac{n-k}{n}\right)^p \tfrac{1}{n}, \\ &\leq 8CM \int_0^1 x(1-x)^p dx = \frac{8CM}{(p+1)(p+2)}; \end{aligned}$$

hence for p so large that $\dfrac{8C}{(p+1)(p+2)} \leq 1$ we have

$$|\vec{\xi}_n| \leq Mn^p.$$

The Dirichlet expansion converges uniformly and the limit V is analytic in both $x > 0$ and B; from classic results [8], Y_1 has an analytic extension to the whole line. The asymptotic behavior of Y_1 as $x \to \infty$ follows readily from the form of the Dirichlet expansion. For $B = 0$ an induction argument shows that the formula (3.8) for $\vec{\xi}_n$ is solved by

$$\vec{\xi}_n = \frac{2(-1)^n}{1+\mu_1} \vec{\xi}_0 \qquad \text{for } n = 1, 2, 3, \ldots,$$

and summing the Dirichlet expansion in closed form yields the solution (3.5).

4. The Evans function. The Evans function detects the intersection of the stable and unstable manifolds of (3.1). From the Dirichlet expansion, Theorem 3.2, and the symmetries of Lemma 3.1, we have explicitly constructed the functions $Y_j = (y_{j1}, y_{j2}, y_{j3}, y_{j4})^T$, which span these manifolds. Placed as columns in a matrix $\mathcal{Y} = [Y_1, Y_2, Y_3, Y_4]$ they form a matrix solution of the first order system (3.1). The Evans function can be written in several equivalent forms; see [4, 30, 15, 16] and the references therein. In the case at hand the trace of the matrix $A(\lambda)$ is zero, and the Evans function reduces to the following normalized determinate of $\mathcal{Y}$, which depends upon λ, B, and ρ but not upon x :

$$E(\lambda, B; \rho) = \frac{-1}{16\mu_1\mu_2}|\mathcal{Y}|. \tag{4.1}$$

The properties of the Evans function are given in the following theorem.

THEOREM 4.1. *The Evans function is analytic for all values of* B, $\rho \in (0,1)$, *and all* $\lambda \notin \{z|\Re z \leq 0 \text{ and } \Im z = \pm\frac{\rho^2}{2}\}$. *For fixed* B, *the zeros* λ *of the Evans function which lie in* $\Omega = \{z|\Re z > 0 \text{ or } |\Im z| < \frac{\rho^2}{2}\}$ *coincide with the eigenvalues of the linear operator* L_2, *and the order of such a zero equals the algebraic multiplicity of the eigenvalue of* L_2. *For all* $\lambda \in \overline{\Omega}$, $E(\lambda^*) = E(\lambda)^*$, *and* $E(\lambda) = E(-\lambda)$ *if* $-\lambda \in \Omega$. *As* $|\lambda| \to \infty, E(\lambda) \to 1$.

Proof. The analyticity of the Evans function follows from that of the constituent solutions Y_j afforded by Theorem 3.2, and from the choice of branch cuts of $\mu_j(\lambda)$. The analytic extension of the Evans function through the essential spectrum is transparent from the Dirichlet expansion construction; there is no need for an application of the Gap lemma; see [11, 18]. The coincidence of zeros of the Evans function in Ω and the eigenvalues of the associated operator, up to multiplicity, is a well known result which can be found in [4]. Since L_2 is a real operator, $L_2V = \lambda V$ if and only if $L_2V^* = \lambda^*V^*$, and the symmetry of the Evans function, for λ in Ω, follows. The asymptotic property of E as $|\lambda| \to \infty$ is a consequence of the normalization and Corollary 1.18 of [30]. □

Remark. The zeros of the extended Evans function which lie outside the domain Ω do not necessarily correspond to eigenvalues of L_2. Also, the symmetries $E(\lambda^*) = E(\lambda)^*$ and $E(\lambda) = E(-\lambda)$ do *not* hold for λ outside of $\overline{\Omega}$.

In that which follows we will view ρ as fixed, and write $E = E(\lambda, B)$. For $B = 0$ the matrix solution $\mathcal{Y}$ takes the simple form

$$\mathcal{Y} = \begin{pmatrix} y_1 & y_2 & y_3 & y_4 \\ iy_1 & -iy_2 & iy_3 & -iy_4 \\ y_{1x} & y_{2x} & y_{3x} & y_{4x} \\ iy_{1x} & -iy_{2x} & iy_{3x} & -iy_{4x} \end{pmatrix}, \tag{4.2}$$

where y_j are given explicitly by (3.5) and the symmetries of Lemma 3.1. We may directly evaluate the Wronskian in (4.2) to obtain the following corollary.

COROLLARY 4.2. *For* $B = 0$ *the Evans function is given by*

$$E = \frac{(1-\mu_1)(1-\mu_2)}{(1+\mu_1)(1+\mu_2)}. \tag{4.3}$$

The operator L_2 *has exactly two eigenvalues* $\lambda_\pm = \pm i\frac{1-\rho^2}{2}$ *which correspond to the localized eigenvectors given by* (3.5).

Proof. We use the explicit formula (3.5) for the functions $Y_j, j = 1, \dots, 4$. The Evans function has exactly two zeros, which lie on the boundary of Ω. As such they may not correspond to localized eigenvalues, but for $\lambda = \lambda_\pm$ inspection of the formula (3.5) shows that the eigenvector is localized in space. □

Each of the eigenvalues $\lambda_\pm$ moves with changes in B. From the symmetry of the spectrum of L_2, it is sufficient to study the path of the eigenvalues bifurcating from λ_+. We will focus on the eigenvalue in the closed first quadrant of the complex plain, denoting it as $\lambda(B)=\lambda(B,\rho)$. Recall that $\lambda(0,\rho)=\lambda_+$ is embedded in the essential spectrum of L_2 for $0<\rho\leq\frac{1}{\sqrt{2}}$. We calculate the derivatives of $\lambda(B)$, for fixed ρ, via the implicit function theorem. First we evaluate the partial derivatives of the Evans function at $\lambda=\lambda_+$. Let $'$ denote differentiation with respect to x and a subscript B denote differentiation with respect to B. Each function Y_j satisfies (3.1), taking the B derivative

$$Y'_{jB}=AY_{jB}+A_BY_j,$$

and since $\mathcal{Y}$ is a matrix solution of the corresponding homogeneous equation, variation of parameters yields

$$Y_{jB}(x)-\mathcal{Y}(x)\mathcal{Y}^{-1}(x_j)Y_{jB}(x_j)=\mathcal{Y}(x)\int_{x_j}^{x}\mathcal{Y}^{-1}(s)A_B(s)Y_j(s)ds.$$

From the Dirichlet expansion formulas we know that $|Y_{jB}(x)|$ decays as $e^{-(\mu_j+2)x}$ as $x\to\infty$ for $j=1,2$, and similarly $|Y_{jB}|$ decays as $x\to-\infty$ for $j=3,4$. We let $x_j\to\infty$ for $j=1,2$ and $x_j\to-\infty$ for $j=3,4$. The second term on the left-hand side converges to zero in each case as $|x|\to\infty$. Defining the matrix Φ as

$$\Phi=\left[\int_{\infty}^{x}\mathcal{Y}^{-1}A_BY_1ds,\ \int_{\infty}^{x}\mathcal{Y}^{-1}A_BY_2ds,\ \int_{-\infty}^{x}\mathcal{Y}^{-1}A_BY_3ds,\ \int_{-\infty}^{x}\mathcal{Y}^{-1}A_BY_4ds\right], \tag{4.4}$$

we can rewrite the four equations as a matrix differential equation in B,

$$\mathcal{Y}_B(x,B)=\mathcal{Y}(x,B)\Phi(x,B). \tag{4.5}$$

Applying Abel's formula [8] we obtain a differential equation for the Wronskian $W=|\mathcal{Y}|$ as

$$W_B=(\operatorname{tr}\Phi)W. \tag{4.6}$$

Since $\operatorname{tr}\Phi'=\operatorname{tr}(\mathcal{Y}^{-1}A_B\mathcal{Y})=\operatorname{tr}A_B=0$ it follows that $\operatorname{tr}\Phi$ is independent of x. We let $x\to\infty$ in (4.4) obtaining

$$\operatorname{tr}\Phi=\operatorname{tr}\int_{-\infty}^{\infty}\left[0,\ 0,\ \mathcal{Y}^{-1}A_BY_3,\ \mathcal{Y}^{-1}A_BY_4\right]dx.$$

Using the notation $y_j=y_{j1}$, we observe that $y_{j2}=(-1)^{j+1}iy_j$ for $j=1,2,3,4$. To invert $\mathcal{Y}$ we expand by cofactors

$$[\mathcal{Y}^{-1}]_{jk}=\frac{(-1)^{j+k}W_{kj}}{W},$$

where W_{kj} is the determinant of the kj-minor of $\mathcal{Y}$. For $B=0$, the minors may be evaluated directly, yielding

$$W_{43}=-4i\beta_2y_1 \text{ and } W_{44}=-4i\beta_1y_2,$$

where $\beta_j=\frac{\mu_j(\mu_j-1)}{1+\mu_j}$ for $j=1,2$. The matrix A_B has all zero entries except for $[A_B]_{4,2}=4\phi^2$. The nonzero diagonal entries of Φ are

$$\Phi_{33}=\int_{-\infty}^{\infty}\frac{-W_{43}}{W}4\phi^2y_{32}ds=-\frac{16\beta_2}{W}\int_{-\infty}^{\infty}\phi^2y_1y_3ds$$

and

$$\Phi_{44} = \int_{-\infty}^{\infty} \frac{W_{44}}{W} 4\phi^2 y_{22} ds = -\frac{16\beta_1}{W} \int_{-\infty}^{\infty} \phi^2 y_2 y_4 ds.$$

Finally we see that

$$\operatorname{tr}\Phi = -\frac{16}{W} \int_{-\infty}^{\infty} \phi^2 \left(\beta_1 y_2 y_4 + \beta_2 y_1 y_3\right) ds = -\frac{32}{W} \left(\beta_1 \gamma_2 + \beta_2 \gamma_1\right),$$

where $\gamma_j = \frac{\mu_j^2 - \frac{1}{3}}{(1+\mu_j)^2}$ for $j = 1, 2$.

Since $E = -\frac{W}{16\mu_1\mu_2}$ we have $E_B(\lambda, 0) = -\frac{W}{16\mu_1\mu_2} \operatorname{tr}\Phi(\lambda, 0)$, and substituting the formulas above,

$$E_B(\lambda, 0) = \frac{2}{\mu_1\mu_2} \left(\beta_1\gamma_2 + \beta_2\gamma_1\right). \tag{4.7}$$

In particular, at the eigenvalue $\lambda_+ = i\frac{1-\rho^2}{2}$, we have $\mu_2 = 1$ and $\mu_1 = i\alpha$ with $\alpha = \sqrt{1-2\rho^2}$, for $0 < \rho < \frac{1}{\sqrt{2}}$; consequently

$$E_B(\lambda_+, 0) = \frac{i\alpha - 1}{3(1+i\alpha)}.$$

Taking the λ partial derivative of (4.3) we calculate

$$E_\lambda(\lambda_+, 0) = \frac{i}{2} \frac{1-i\alpha}{1+i\alpha}.$$

By the implicit function theorem, the function $\lambda(B)$ which satisfies $E(\lambda(B), B) = 0$ and $\lambda(0) = \lambda_+$ has derivative

$$\lambda_B(0) = -\frac{E_B(\lambda_+, 0)}{E_\lambda(\lambda_+, 0)} = -\frac{2}{3}i. \tag{4.8}$$

To leading order in B, the embedded eigenvalue moves along the imaginary axis as B increases from zero. This result is in fact a consequence of the method developed in [16, Theorem 1.1 and Lemma 4.3] to compute the leading order motion of eigenvalues. To determine if the eigenvalue remains in the essential spectrum and on the imaginary axis, it is necessary to compute the second derivative of $\lambda(B)$ with respect to B at $B = 0$. The implicit function theorem yields

$$\lambda_{BB}(0) = -\frac{E_{\lambda\lambda}\lambda_B^2 + 2E_{\lambda B}\lambda_B + E_{BB}}{E_\lambda}.$$

Taking partial derivatives with respect to λ of (4.3) and (4.7) we obtain

$$E_{\lambda\lambda}(\lambda_+, 0) = \frac{2 - i\alpha - i\alpha^3}{i\alpha(i\alpha+1)^2} \tag{4.9}$$

and

$$E_{\lambda B}(\lambda_+, 0) = -\frac{1}{3} \frac{(i\alpha - 1)(i\alpha + 3)}{\alpha(i\alpha + 1)}. \tag{4.10}$$

It remains to calculate $E_{BB}(\lambda_+,0)$. This can be done directly by differentiating (4.6) with respect to B, obtaining $W_{BB}=\left(\operatorname{tr}\Phi_B+(\operatorname{tr}\Phi)^2\right)W$, for $B=0$ and $\lambda\neq\lambda_\pm$, taking the limit $\lambda\to\lambda_+$, and expanding in terms of $\epsilon=\mu_2-1$ as $\mathcal{Y}^{-1}$ becomes singular. However, it is easier to avoid the singularity altogether and compute W_{BB} from the vectors Y_{jB} and Y_{jBB}. Since $y_2=y_4=\frac{1}{2}\mathrm{sech}x$ at $\lambda=\lambda_+$ and $B=0$, after cancelation we obtain

$$\begin{aligned}W_{BB}=&\Big|[Y_1,Y_2,Y_3,Y_{4BB}-Y_{2BB}]\Big|+2\Big|[Y_1,Y_{2B},Y_3,Y_{4B}]\Big|\\&+2\Big|[Y_{1B},Y_2,Y_3,Y_{4B}-Y_{2B}]\Big|+2\Big|[Y_1,Y_2,Y_{3B},Y_{4B}-Y_{2B}]\Big|.\end{aligned}\tag{4.11}$$

The matrix $\mathcal{Y}$ is singular at the eigenvalue and there is a solution of (3.1), Y_5, which is linearly independent of $\{Y_1,Y_2,Y_3\}$. Using reduction of order it is straightforward to find $Y_5=(y_5,-iy_5,y_{5x},-iy_{5x})^T$, where

$$y_5=\frac{1}{2}\mathrm{sech}x\left(x+\frac{\sinh 2x}{2}\right).$$

We form the matrix solution $\mathcal{Y}_0=[Y_1,Y_2,Y_3,Y_5]$, and as before, we solve for $\mathcal{Y}_B$ via variation of parameters at $\lambda=\lambda_+$, viz.

$$\mathcal{Y}_B=\mathcal{Y}_0\Phi_0,\tag{4.12}$$

where

$$\Phi_0=\left[\int_\infty^x\mathcal{Y}_0^{-1}A_BY_1ds,\int_\infty^x\mathcal{Y}_0^{-1}A_BY_2ds,\int_{-\infty}^x\mathcal{Y}_0^{-1}A_BY_3ds,\int_{-\infty}^x\mathcal{Y}_0^{-1}A_BY_4ds\right].\tag{4.13}$$

Similarly, for $j=1,2,3,4$, the second derivatives Y_{jBB} satisfy the system of differential equations

$$Y'_{jBB}=AY_{jBB}+2A_BY_{jB}.$$

Variation of parameters again yields

$$Y_{jBB}=2\mathcal{Y}_0\int_{x_j}^x\mathcal{Y}_0^{-1}A_BY_{jB}ds,\tag{4.14}$$

where $x_j=\infty$ for $j=1,2$ and $x_j=-\infty$ for $j=3,4$. Substituting the expressions from (4.12) and (4.14) into the formula (4.11) leads to, after some computation given in the appendix, the following equality:

$$\Re\left(\frac{E_{BB}(\lambda_+,0)}{E_\lambda(\lambda_+,0)}\right)=\frac{8(2-3\rho^2)}{9\alpha(1-\rho^2)}-\frac{2\alpha(1-\rho^2)\pi^2}{9}\mathrm{sech}^2\frac{\alpha\pi}{2}.\tag{4.15}$$

We separate the other terms in the expression for $\lambda_{BB}(0)$,

$$\Re\left(\frac{2E_{\lambda B}\lambda_B}{E_\lambda}\right)=-\frac{8}{3\alpha}$$

and

$$\Re\left(\frac{E_{\lambda\lambda}\lambda_B^2}{E_\lambda}\right)=\frac{8}{9\alpha(1-\rho^2)}.$$

Combining these results we obtain the leading order expression for the motion of the real part of the embedded eigenvalue with respect to B,

$$\Re\left(\lambda_{BB}(0)\right) = \frac{2\alpha(1-\rho^2)\pi^2}{9}\operatorname{sech}^2\frac{\alpha\pi}{2}. \tag{4.16}$$

We summarize and extend the results above in the following theorem.

THEOREM 4.3. *For $B = 0$, the eigenvalue problem has two simple eigenvalues $\lambda = \pm i\frac{1-\rho^2}{2}$. For fixed ρ satisfying $\rho \in (0, \frac{1}{\sqrt{2}})$, the eigenvalues depend analytically upon B. They are embedded in the essential spectrum for $B = 0$ and leave the imaginary axis as B increases from zero, forming two complex pairs of eigenvalues. To leading order in real and imaginary parts, the unique eigenvalue in the open first quadrant of the complex plans satisfies*

$$\begin{aligned}\lambda(B) = \quad & i\left(\tfrac{1-\rho^2}{2} - \tfrac{2}{3}B + O(B^2)\right)\\ & + \left(\tfrac{2\pi^2(1-\rho^2)\sqrt{1-2\rho^2}}{9}\operatorname{sech}^2\left(\tfrac{\pi\sqrt{1-2\rho^2}}{2}\right)\right)B^2 + O(B^3).\end{aligned} \tag{4.17}$$

In particular the up wave is structurally unstable for $\rho \in (0, \frac{1}{\sqrt{2}})$. For $B = 0$ and $\rho \in (\frac{1}{\sqrt{2}}, 1)$, the eigenvalues are simple and isolated and remain upon the imaginary axis for B small enough; moreover, the unique eigenvalue of positive imaginary part satisfies

$$\lambda(B) = i\left(\frac{1-\rho^2}{2} - \frac{2}{3}B + O(B^2)\right). \tag{4.18}$$

Proof. The asymptotic formula for B small and $\rho \in (0, \frac{1}{\sqrt{2}})$ follow from the preceding discussion. The uniqueness of the eigenvalue in the open first quadrant follows from the fact that there is at most one set of four strictly complex eigenvalues, as observed in section 2. For $\rho \in (\frac{1}{\sqrt{2}}, 1)$, Corollary 4.2 indicates that the eigenvalues are isolated and simple. Since the spectrum is symmetric with respect to the imaginary axis, the eigenvalues can only leave the imaginary axis in pairs, which necessitates collision with another eigenvalue, or the essential spectrum. Thus the eigenvalues are trapped upon the imaginary axis for B small enough, all the derivatives of λ with respect to B have zero real part at $B = 0$, and the formula (4.8) holds. □

We now address the case when the embedded eigenvalues $\lambda_\pm = \pm i\frac{1-\rho^2}{2}$ of the operator L_2 coincide with the branch points $\pm i\frac{\rho^2}{2}$; i.e., $\rho = \frac{1}{\sqrt{2}}$. In this case the Evans function is not analytic at the eigenvalue, and we must unfold the Riemann surface to investigate the behavior of its zeros as B is perturbed from zero. Indeed, the Evans function in the new independent variable $\nu = \sqrt{\rho^2 + 2i\lambda}$ has a simple pole in ν at $\nu = 0$, which we eliminate and study the zeros of the function

$$D(\nu, B) = 16\nu\sqrt{2\rho^2 - \nu^2}E\left(i\frac{\rho^2 - \nu^2}{2}, B\right) \tag{4.19}$$

in a neighborhood of $\nu = 0$. Note that when $-\frac{\pi}{4} < \arg\nu < \frac{3\pi}{4}$, there is a unique λ with $-\pi < \arg(\lambda - i\rho^2/2) < \pi$ such that $\nu = \sqrt{\rho^2 + 2i\lambda}$. In particular, a zero $\nu = \nu_0$ of D corresponds to a zero $\lambda_0 = i\frac{\rho^2 - \nu_0^2}{2}$ of E if and only if $-\frac{\pi}{4} < \arg\nu_0 < \frac{3\pi}{4}$. A zero of D which does not satisfy this condition will be termed *spurious.*

Lemma 4.4. *In the new independent variable $\nu = \sqrt{\rho^2 + 2i\lambda}$, there is a neighborhood of the point $(\nu = 0, \rho = \frac{1}{\sqrt{2}})$ for which the function D is analytic with respect to ν, ρ, and B.*

Proof. The function D defined above is the Wronskian (4.2); its analyticity follows directly from that of the functions $Z_j(x, \nu, \rho, B) = Y_j(x, i\frac{\rho^2-\nu^2}{2}, B)$ for $j = 1, \dots, 4$. Each Z_j is given as a Dirichlet expansion, as in Theorem 1 for Y_j, whose terms are analytic with respect to ν, ρ, and B when $|\nu|^2 < 1 - 2\delta$ and $|\rho - \frac{1}{\sqrt{2}}| < \delta$ for any fixed constant δ with $0 < \delta < 1/4$. From the uniform convergence of the derivatives of the Dirichlet expansion with respect to these parameters, we conclude that Z_j is analytic whenever the expansion is defined, and the analytic expansion of the derivatives to the entire x-axis leads to the conclusion. □

When $\rho = \frac{1}{\sqrt{2}}$, the function

$$D(\nu, B) = 16\nu\sqrt{1-\nu^2}E\left(i\frac{\rho^2 - \nu^2}{2}, B\right) = -|\mathcal{Y}|_{\lambda = i\frac{1/2-\nu^2}{2}} \tag{4.20}$$

is analytic with respect to ν and B in the open set $\{|\nu| < 1\} \times \{|B| < \infty\}$, and from formula (4.3) of Corollary 4.2 we obtain the expression

$$D(\nu, 0) = \frac{16\nu^3(1-\nu^2)^{3/2}}{(1+\sqrt{1-\nu^2})^2(1+\nu)^2}. \tag{4.21}$$

In particular, D has a zero of the multiplicity three at $\nu = 0$ for $\rho = \frac{1}{\sqrt{2}}$. When $B > 0$ three zeros move away from the origin, but only one corresponds to a zero of the Evans function.

Theorem 4.5. *Let $\rho = \frac{1}{\sqrt{2}}$. There exists an $\epsilon > 0$ such that for all B satisfying $0 < B < \epsilon$, the operator L_2 has exactly one pair $\pm\lambda$ of eigenvalues where λ lies on the imaginary axis between 0 and $\frac{i}{4}$, and satisfies*

$$\lambda(B) = i\left(\frac{1}{4} - \frac{2}{3}B + o(B)\right). \tag{4.22}$$

Proof. On the circle $|\nu| = \frac{1}{\sqrt{2}}$ we may use the formula (4.21) to obtain the lower bound

$$|D(\nu, 0)| \geq \frac{1}{8};$$

thus for $|B|$ sufficiently small, $|D(\nu, B)| > 0$ on $|\nu| = 1/\sqrt{2}$, and since D is analytic in $V_1 = \{|\nu| \leq \frac{1}{\sqrt{2}}\}$ the number of zeros of D in V_1 is constantly 3 for B small enough. To follow the motion of these zeros with respect to B we expand D as

$$D(\nu, B) = D(\nu, 0) + D_B(\nu, 0)B + g(\nu, B)B^2,$$

where $g(\nu, B)$ is an analytic function with respect to both B and ν. Rewriting (4.7) in the new variable ν, we find from (4.20) that

$$D_B(\nu, 0) = \frac{32\nu\sqrt{1-\nu^2}\left[\nu(1/3 - \nu^2) + \sqrt{1-\nu^2}\,(\nu^2 - 2/3)\right]}{(1+\nu)^2(1+\sqrt{1-\nu^2}\,)^2} = -\frac{16}{3}\nu + O(\nu^2). \tag{4.23}$$

The function $f(\nu, B) = D(\nu, 0) + D_B(\nu, 0)B$ has, by arguments similar to those above, exactly three zeros in V_1 for B small enough. Writing

$$f(\nu, B) = D_B(\nu, 0)\left[B + \frac{D(\nu, 0)}{D_B(\nu, 0)}\right],$$

where

$$\frac{D(\nu, 0)}{D_B(\nu, 0)} = \frac{\nu^2(1-\nu^2)}{2[\nu(1/3-\nu^2) + \sqrt{1-\nu^2}\,(\nu^2 - 2/3)]} = -\frac{3}{4}\nu^2 + O(\nu^3), \tag{4.24}$$

it is easy to see that $B + \frac{D(\nu,0)}{D_B(\nu,0)}$ has two real zeros, $\nu_{-1} = -\sqrt{\frac{4}{3}B} + O(B)$ and $\nu_1 = \sqrt{\frac{4}{3}B} + O(B)$ for $B > 0$ and small enough, while $\nu_0 = 0$ is always a zero of f.

To localize the zeros of D we apply Rouche's theorem. Fixing $\alpha > 0$ small we show that the bound $|f| > B^2|g|$ holds on each of the two circles $\mathcal{U}_{\pm 1} = \{\nu : |\nu - \nu_{\pm 1}| = |\nu_{\pm 1}|^{2-\alpha}\}$. We examine only the case $\nu = \nu_1$. Substituting $B = -\frac{D(\nu_1,0)}{D_B(\nu_1,0)} = -\frac{3}{4}\nu_1^2 + O(\nu_1^3)$ into f and writing $\nu = \nu_1 + z$ where $|z| = |\nu_1|^{2-\alpha}$, computation leads to the asymptotic relation

$$\begin{aligned} |f(\nu_1 + z, B)| &= |D_B(\nu_1, 0)| \left| -\frac{D(\nu_1, 0)}{D_B(\nu_1, 0)} + \frac{D(\nu_1 + z, 0)}{D_B(\nu_1 + z, 0)} \right| \\ &= \tfrac{16}{3}|\nu_1| \left| -\tfrac{3}{2}\nu_1 z - \tfrac{3}{4}z^2 \right| + O(|\nu_1|^3, |z|^2|\nu_1|), \end{aligned}$$

from which we deduce $|f| > c_1|\nu_1|^{4-\alpha}$ for some positive constant c_1 and for $B > 0$ small enough. However, since $|B| = O(|\nu_1|^2)$ it follows that $|B^2 g| < c_2|\nu_1|^4$ for some positive c_2 and small $B > 0$. We deduce from Rouche's theorem that $D = f + B^2 g$ has exactly one zero inside $\mathcal{U}_{\pm 1}$ for $B > 0$ small enough. The circle $\mathcal{U}_{-1}$ lies inside the spurious region and zero of D inside $\mathcal{U}_{-1}$ is spurious. On the other hand, all $\nu \in \mathcal{U}_1$ satisfy $-\frac{\pi}{4} < \arg\nu < \frac{3\pi}{4}$ and hence $\nu = \sqrt{\frac{1}{2} + 2i\lambda}$ for some λ in a neighborhood of the branch point $\frac{i}{4}$. The zero of D in $\mathcal{U}_1$ corresponds to a zero of E.

The localization of the third, smallest zero of D requires a slightly different analysis. Since $D(0,0) = D_B(0,0) = D_\nu(0,0) = D_{\nu\nu}(0,0) = 0$, the Taylor expansion of D about $(0,0)$ takes the form

$$D(\nu, B) = (\nu B)D_{\nu B}(0,0) + \left(\frac{1}{2}B^2\right) D_{BB}(0,0) + R(\nu, B), \tag{4.25}$$

where R is analytic and satisfies $|R(\nu, B)| \le K_1(|B|^3 + |\nu|^3)$ for some constant $K_1 > 0$ and B and ν small enough. We may calculate directly from (4.19) that

$$D_{\nu B}(0,0) = 16E_B\left(\frac{i}{4}, 0\right) - \lim_{\nu \to 0} 16i\nu^2\sqrt{1-\nu^2}E_{B\lambda}\left(i\frac{\frac{1}{2} - \nu^2}{2}, 0\right),$$

and from (4.7) we easily deduce that the limit term is zero while $E_B(\frac{i}{4}, 0) = -\frac{1}{3}$, yielding $D_{\nu B}(0,0) = -\frac{16}{3}$. Since D is analytic in ρ we may compute from (6.5)

$$D_{BB}(0,0) = \lim_{\rho \to \frac{1}{\sqrt{2}}} 16\mu_1(\lambda_+(\rho))\mu_2(\lambda_+(\rho))E_{BB}\left(i\frac{1-\rho^2}{2}, 0\right) = -\frac{64}{9}.$$

Defining

$$f_0(\nu, B) = -\frac{16}{3}B\left(\nu + \frac{4}{3}B\right),$$

we have $D(\nu, B) = f_0(\nu, B) + R(\nu, B)$, where f_0 has a zero at $\nu_0 = -\frac{4}{3}B$. On the ball $|\nu - \nu_0| = \frac{2}{3}|B|$, we have $|f_0(\nu, B)| = \frac{64}{27}B^2$, but $|R(\nu, B)| \leq K_1(B^3 + |\nu|^3) \leq K_1(B^3 + \frac{5}{3}B^3)$. It follows from Rouche's theorem that for B small enough, D has exactly one zero inside $\mathcal{U}_0 = \{\nu : |\nu - \nu_0| = \frac{2}{3}|B|\}$. In particular if $B > 0$ this zero is spurious and does not correspond to a zero of the Evans function E.

We have shown that for $B > 0$, but small enough, D has exactly one zero in V_1, $\nu_1 = \sqrt{\frac{4}{3}B} + O(B)$ satisfying $-\frac{\pi}{4} < \arg \nu_1 < \frac{3\pi}{4}$. Indeed we show that this zero lies on the real axis. From Lemma 4.4, $D(\nu, B) = 16\nu\sqrt{1-\nu^2}E(i\frac{1/2-\nu^2}{2}, B)$ is a real valued function when $0 \leq \nu \leq 1$. The equalities $D(0, 0) = D_B(0, 0) = 0$ and $D_{BB}(0, 0) = -\frac{64}{9}$ imply that $D(0, B) < 0$ when $B > 0$ is sufficiently small. In addition, since $D(\frac{1}{\sqrt{2}}, 0) > 0$ and D is continuous in B, it follows that $D(\frac{1}{\sqrt{2}}, B) > 0$ for any $B > 0$ sufficiently small. Therefore, D has at least one zero for ν in the interval $(0, \frac{1}{\sqrt{2}})$. But D has only three zeros in V_1, two of which do not lie on $(0, \frac{1}{\sqrt{2}})$, and from the analysis above this zero must be ν_1. The asymptotic formula (4.22) for $\lambda(B)$ follows from the relation $\lambda(B) = i\frac{\frac{1}{2}-\nu_1^2}{2}$ and the expansion for ν_1. □

While the case $B < 0$ is not directly relevant to the birefringence modeled by the coupled nonlinear Schrödinger equations studied here, it has mathematical interest. Recalling the new variable $\nu = \sqrt{\rho^2 + 2i\lambda}$, the point $\lambda = 0$ corresponds to $\nu = \frac{1}{\sqrt{2}}$ and as observed above $D(\frac{1}{\sqrt{2}}, B) > 0$ for $|B|$ sufficiently small, while $D(0, 0) = 0$. Moreover, D is real when $\nu \in [0, \frac{1}{\sqrt{2}}]$, and whenever $D(0, B)$ is negative for small $|B|$, there will be a zero of D in $(0, \frac{1}{\sqrt{2}})$ which corresponds to an imaginary zero of E between the branch point and the origin. For $\rho = \frac{1}{\sqrt{2}}$ we have shown that

$$D(0, B) = D_{BB}(0, 0)\frac{B^2}{2} + O(|B|^3) = -\frac{32B^2}{9} + O(|B|^3),$$

and for $\rho \neq \frac{1}{\sqrt{2}}$ we have

$$D(0, B) = D_B(0, 0)B + O(B^2) = \frac{32\sqrt{2}\rho(1-\sqrt{2}\rho)}{3(1+\sqrt{2}\rho)^2}B + O(B^2).$$

Clearly for $\rho \in (0, \frac{1}{\sqrt{2}}]$, $D(0, B) < 0$ for $B < 0$ small enough, and as B decreases from 0, the branch points $\pm i\frac{\rho^2}{2}$ shed eigenvalues which travel along the imaginary axis toward the origin. We state this result in the following proposition.

PROPOSITION 4.6. *For $\rho \in (0, \frac{1}{\sqrt{2}}]$, as B decreases from zero, the branch points $\pm i\frac{\rho^2}{2}$ of the essential spectrum of L_2 exhibit an edge bifurcation, each shedding an eigenvalue along the imaginary axis towards the origin.*

5. Numerical computation of the stability diagram. In the far from integrable case, B not small, we no longer have explicit forms for the solutions $Y_j, j = 1, \ldots, 4$, of the linearized system (3.1). The Dirichlet expansions for the Y_j are unwieldy for an exact analysis of the eigenvalue problem. For B not small we turn to a numerical evaluation of the Evans function; however, the asymptotic boundary conditions (3.3) of the special solutions render their direct numerical computation inefficient. The Dirichlet expansion provides a fast and rigorous mechanism for their computation.

From the symmetries of the Y_j given in Lemma 3.1, the Evans function takes a particularly simple form when evaluated at $x = 0$. Recalling the notation $Y_1 = (v_1, v_2, v_{1x}, v_{2x})^T$, the following result holds.

LEMMA 5.1. *The Evans function may be expressed as*

$$E = -\frac{1}{4\mu_1\mu_2}\left(v_1(0,\lambda,B)v_1(0,-\lambda,B) + v_1(0,-\lambda,B)v_2(0,\lambda,B)\right) \times \left(v_{1x}(0,\lambda,B)v_{2x}(0,-\lambda,B) + v_{1x}(0,-\lambda,B)v_{2x}(0,\lambda,B)\right). \tag{5.1}$$

Proof. At $x = 0$ we have $Y_3(0,\lambda,B) = S_1Y_1(0,\lambda,B)$ and $Y_4(0,\lambda,B) = S_1Y_2(0,\lambda,B)$, while $Y_2(0,\lambda,B) = S_2Y_1(0,-\lambda,B)$. Evaluating the determinate leads to the formula above. □

From the uniform convergence for the Dirichlet expansion it may be efficiently summed for values of $x = x_0 > 0$. We may then take $Y_1(x_0,\lambda,B)$ as initial data for the linear system (3.1) and numerically integrate it from $x \in [0, x_0]$, an interval upon which the system is not stiff, thereby obtaining an accurate and efficient approximation to $Y_1(0,\lambda,B)$. In particular, it is straightforward to rigorously bound the numerical error in such an approximation.

We present the numerical results obtained by fixing $\rho \in (0,1)$ and increasing B from 0 to $1.05B_c(\rho)$, where at $B = B_c(\rho) = \frac{2-\rho-\rho^2}{4}$ there is an eigenvalue at 0 whose multiplicity is either two or four. We find five distinct cases:

(1) $\rho \in (0, \rho_c)$ where $\rho_c \approx 0.3242\ldots$,
(2) $\rho = \rho_c$,
(3) $\rho \in (\rho_c, \rho_b)$ where $\rho_b = \frac{1}{\sqrt{2}}$,
(4) $\rho = \rho_b$,
(5) $\rho \in (\rho_b, 1)$.

Of the four eigenvalues of the point spectrum, we denote by $\lambda_1(B,\rho)$ the eigenvalue of largest real part, and when two have the same real part, the one with smallest positive imaginary part. An eigenvalue which is not a symmetry of λ_1 will be denoted λ_2. When λ_1 is purely real or imaginary it is also the eigenvalue of L_2 of negative Krein sign.

For case (1) (see Figure 1(a)) with $B = 0$, the eigenvalue $\lambda_1(0,\rho) = \lambda_+ = i\frac{1-\rho^2}{2}$ is embedded in the essential spectrum and by Theorem 4.3 it leaves the imaginary axis quadratically in B. From the numerical computation we find that as B increases the eigenvalue moves toward the real axis, forming a Jordan pair on the real axis with its complex conjugate when $B = B_1(\rho) < B_c(\rho)$. As B increases further the pair split with λ_2 moving toward and reaching the origin at $B = B_c$, and λ_1 moving outward from the origin along the real axis with increasing B. When λ_2 reaches the origin, it combines with its negative, and this pair splits into complex conjugates which move outward along the imaginary axis until they vanish at the branch points of the essential spectrum.

In the critical case, $\rho_c \approx 0.3242\ldots$ (see Figure 1(b)), the path traced by λ_1 and its symmetries "pinches" at the origin. Specifically, the embedded eigenvalue leaves the essential spectrum quadratically in B as it increases from $B = 0$, moving through the complex plane toward the origin, where the four recombine, forming a Jordan chain of multiplicity four at the origin as $B = B_c(\rho_c)$. For $B > B_c$, λ_1 and its negative move outward from the origin along the real axis and λ_2 and its negative move outward along the imaginary axis until they hit the branch point at $\pm i\frac{\rho^2}{2}$, becoming spurious.

For case (3), the complex eigenvalues ejected from the essential spectrum at $B = 0$

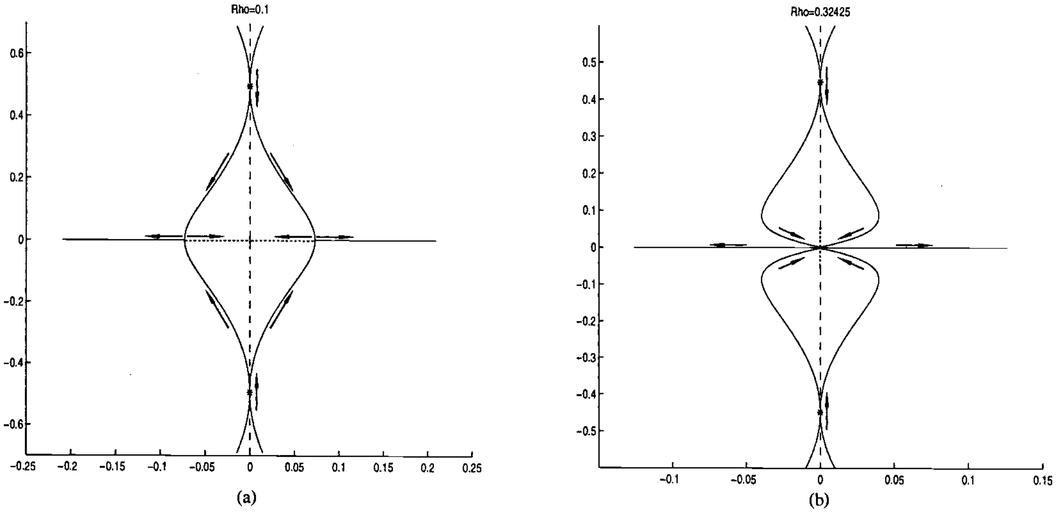


FIG. 1. *The spectrum of* $L_2(\rho, B)$ *for* (a) $\rho = 0.1$ *and* (b) $\rho = 0.32425$. *Solid line: location of negative Krein sign eigenvalue; dashed line: essential spectrum; dot: branch point; dotted line: positive Krein sign eigenvalue; arrows: direction of motion with increasing* B. *The* $*$ *indicates location of eigenvalue at* $B = 0$. *In both figures,* B *ranges from* $-0.75B_c(\rho)$ *to* $1.05B_c(\rho)$. *In figure* (a) *the branch points almost touch the real axis. In figure* (b) *the four complex eigenvalues meet at the origin.*

land upon the exposed neck of the imaginary axis; viewing Figures 1(a)–(b) and 2(a)–(b) sequentially, as ρ increases, the essential spectrum recedes from the origin until at the critical value the corresponding path of the complex eigenvalues pinches down to a point at the origin, and for values of $\rho > \rho_c$ the path of the complex eigenvalues "pulls up" on the imaginary axis. It is this sequence of events which we term a *neck bifurcation.* Figure 2(a) shows that as B increases from $B = 0$, the embedded eigenvalue λ_1 leaves the imaginary axis quadratically in B, and at $B = B_2(\rho) < B_c(\rho)$ it returns to the imaginary axis, forming a Jordan pair with $-\lambda_1^*$ at some point below the branch point $i\frac{\rho^2}{2}$. This pair splits, λ_1 moves down the imaginary axis, forming a Jordan pair at the origin with its complex conjugate for $B = B_c(\rho)$, and λ_2 moves up the imaginary axis, hits the branch point, and becomes spurious. This motion of λ_2 is not shown on Figure 2 due to smallness of scale. For $B > B_c$ the Jordan pair at the origin splits, forming a pair $\pm\lambda_1$, which move outward from the origin along the real axis.

For case (4), with $\rho = \frac{1}{\sqrt{2}}$ pictured in Figure 2(b), the eigenvalue λ_1 coincides with the branch point $\frac{i}{4}$ as $B = 0$, arriving there from off the imaginary axis. For clarity of presentation the eigenvalues on the imaginary axis for $B < 0$ are not shown. For $B > 0$ the eigenvalue λ_1 travels along the imaginary axis toward the origin, arriving there at $B = B_c(\rho_b)$ where it forms a Jordan pair with its complex conjugate, splitting into a real pair $\pm\lambda_1$ as $B > B_c(\rho_b)$. For case (5), not pictured, when $B = 0$ the eigenvalue λ_1 is on the imaginary axis between the branch point of the essential spectrum and the origin; as B increases it moves along the imaginary axis toward the origin, forming a Jordan pair with its complex conjugate at the origin as $B = B_c(\rho)$, again splitting into a real pair as B increases further. In case (5) there are only two eigenvalues for $B > 0$.

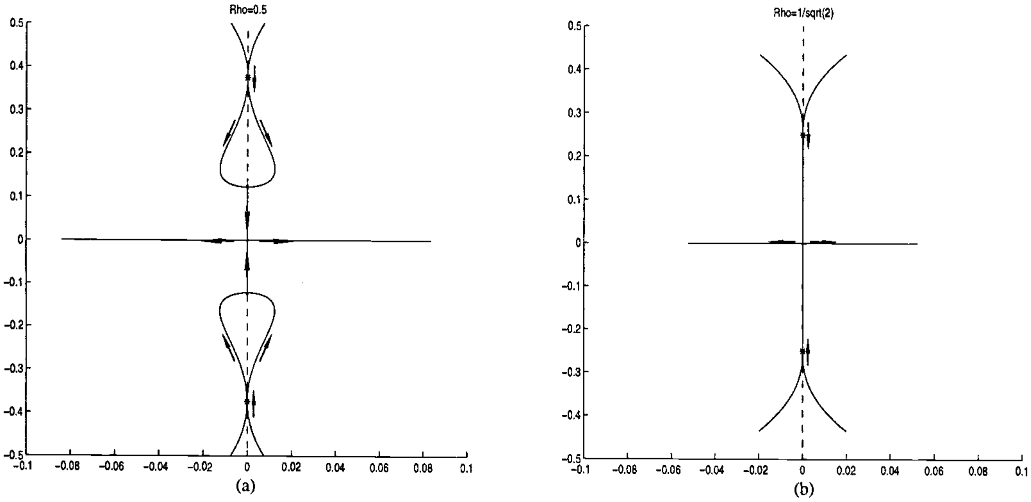


FIG. 2. *The spectrum of* $L_2(\rho, B)$ *for* (a) $\rho = 0.5$ *and* (b) $\rho = \frac{1}{\sqrt{2}}$. *The legend is the same as Figure* 1. B *varies from* $-0.75B_c$ *to* $1.05B_c$. *In* (b) *the embedded eigenvalues coincide with the branch points for* $B = 0$.

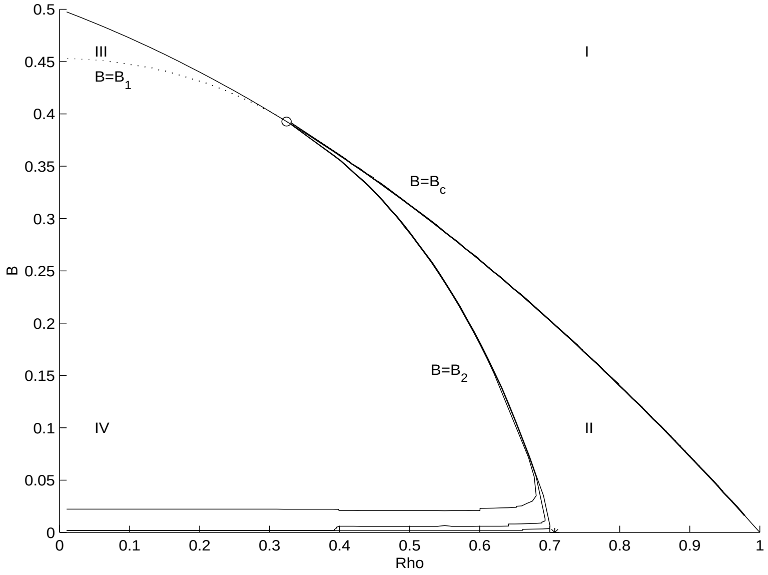


FIG. 3. *The stability diagram of the eigenvalue problem, showing the line* $B = B_c(\rho)$ *(solid), and contours of* $\Re\lambda_1$ *(solid) and* $\Im\lambda_1$ *(dotted) at* $10^{-4}, 10^{-6}$, *and* 10^{-7}.

In Figure 3 we plot the stability diagram comprised of contours of $\Re\lambda_+(B, \rho)$ and $\Im\lambda_+(B, \rho)$ for $(\rho, B) \in [0, 1] \times [0, 1]$. The curves $B = B_c(\rho)$, $B = B_1(\rho)$, and $B = B_2(\rho)$ meet at a common point, indicated with a circle, and divide the square into four regions, labeled I-IV. In region I the eigenvalue λ_1 is real and positive. In region II it is on the imaginary neck between the branch points of the essential spectrum.

This is the region of structural stability of the operator $L_2(\rho, B)$. In region III both λ_1 and λ_2 are real. In the remainder of the bifurcation state space, region IV, there is a set of four complex eigenvalues, symmetric about the real and imaginary axis. The point indicated with the circle in Figure 3 identifies the values $\rho = \rho_c, B = B_c(\rho_c)$ at which the kernel of L_2 has dimension four.

6. Appendix. We calculate here the result given in (4.15). Recalling the definition (4.13) of Φ_0, we substitute the values of A_B and Y_j from (4.2) to obtain

$$\Phi_0 = 4i\left[\int_{\infty}^{x}\phi^2 y_1\mathcal{Y}_0^{-1}\vec{e}_4, \int_{\infty}^{x}-\phi^2 y_2\mathcal{Y}_0^{-1}\vec{e}_4, \int_{-\infty}^{x}\phi^2 y_3\mathcal{Y}_0^{-1}\vec{e}_4, \int_{-\infty}^{x}-\phi^2 y_4\mathcal{Y}_0^{-1}\vec{e}_4\right],$$

where $\vec{e}_4 = (0,0,0,1)^T$. The determinate $W_0 = |\mathcal{Y}_0| = \frac{4i\alpha(1-i\alpha)}{1+i\alpha}$ and from the cofactors of $\mathcal{Y}_0$ we calculate

$$\mathcal{Y}_0^{-1}\vec{e}_4 = \frac{1}{W_0}(-iy_3, -iW_0y_5, iy_1, iW_0y_2)^T.$$

For notational convenience we write Φ_0 in terms of its column vectors as $\Phi_0 = [\vec{\gamma}, \vec{\delta}, \vec{\sigma}, \vec{\nu}]$, and consequently (4.12) implies $Y_{1B} = \mathcal{Y}_0\vec{\gamma}$, $Y_{2B} = \mathcal{Y}_0\vec{\delta}, \ldots$, etc. The second derivative terms (4.14) have a similar form; specifically we denote $\vec{\alpha} = \int_{\infty}^{x}\mathcal{Y}_0^{-1}A_B\mathcal{Y}_0\vec{\delta}ds$ and $\vec{\beta} = \int_{-\infty}^{x}\mathcal{Y}_0^{-1}A_B\mathcal{Y}_0\vec{\nu}ds$, and hence $Y_{2BB} = 2\mathcal{Y}_0\vec{\alpha}$ and $Y_{4BB} = 2\mathcal{Y}_0\vec{\beta}$. In this notation we may expand the determinates in (4.11), finding many zero terms. After simplification we obtain

$$\Big|[Y_1, Y_2, Y_3, Y_{4BB} - Y_{2BB}]\Big| = 2(\beta_4 - \alpha_4)W_0,$$

$$\Big|[Y_1, Y_{2B}, Y_3, Y_{4B}]\Big| = (\delta_2\nu_4 - \delta_4\nu_2)W_0,$$

$$\Big|[Y_{1B}, Y_2, Y_3, Y_{4B} - Y_{2B}]\Big| = \big(\gamma_1(\nu_4 - \delta_4) - \gamma_4(\nu_1 - \delta_1)\big)W_0,$$

and

$$\Big|[Y_1, Y_2, Y_{3B}, Y_{4B} - Y_{2B}]\Big| = \big(\sigma_3(\nu_4 - \delta_4) - \sigma_4(\nu_3 - \delta_3)\big)W_0.$$

The Wronskian W is independent of x and the equalities

$$\text{(6.1)}\ \vec{\delta}(x) = \begin{pmatrix}\delta_1(x)\\ \delta_2(x)\\ \delta_3(x)\\ \delta_4(x)\end{pmatrix} = \begin{pmatrix}\nu_3(-x)\\ \nu_2(-x)\\ \nu_1(-x)\\ -\nu_4(-x)\end{pmatrix}, \qquad \vec{\gamma}(x) = \begin{pmatrix}\gamma_1(x)\\ \gamma_2(x)\\ \gamma_3(x)\\ \gamma_4(x)\end{pmatrix} = \begin{pmatrix}\sigma_3(-x)\\ \sigma_2(-x)\\ \sigma_1(-x)\\ -\sigma_4(-x)\end{pmatrix}$$

hold for any $x \in \mathbb{R}$; we use them to simplify (4.11) as

$$W_{BB} = 2W_0\ \ (\beta_4 - \alpha_4 - 2\delta_2\delta_4 - 4\gamma_1\delta_4 + 2\gamma_4(\delta_1 - \delta_3))\big|_{x=0}.$$

Recalling the equality $E_{BB} = \frac{i}{16\alpha}W_{BB}$ and the relation $E_\lambda = \frac{W_0}{8\alpha}$ valid at $\lambda = \lambda_+$ and $B = 0$, we arrive at the expression

$$\text{(6.2)}\qquad \frac{E_{BB}}{E_\lambda} = i\ \ (\beta_4 - \alpha_4 - 2\delta_2\delta_4 - 4\gamma_1\delta_4 + 2\gamma_4(\delta_1 - \delta_3))\big|_{x=0}.$$

From the definition of $\vec{\delta}$ and $\vec{\gamma}$ we have

$$-2\delta_2(0)\delta_4(0) = 32\int_0^\infty \phi^2 y_2 y_5\,dx \int_0^\infty \phi^2 y_2^2\,dx = \frac{4}{9}(1+2\ln 2),$$

$$-4\gamma_1(0)\delta_4(0) = -\frac{64}{W_0}\int_0^\infty \phi^2 y_1 y_3\,dx \int_0^\infty \phi^2 y_2^2\,dx = -i\frac{8(2-3\rho^2)}{9\alpha\,(1-\rho^2)}, \tag{6.3}$$

and

$$2\gamma_4(0)(\delta_1(0)-\delta_3(0)) = \frac{32}{W_0}\int_0^\infty \phi^2 y_1 y_2\,dx \int_0^\infty \phi^2 (y_1+y_3)\,dx$$

$$= \frac{2\pi \operatorname{sech}\frac{\alpha\pi}{2}}{9}\left[1+2\alpha\int_0^\infty \operatorname{sech}^3(x)\,\sin(\alpha\,x)\,dx + i(\alpha(1-\rho^2)\pi)\operatorname{sech}\frac{\alpha\pi}{2}\right].$$

It remains to evaluate the quantity $\beta_4(0)-\alpha_4(0)$. From the definition of $\vec{\alpha}$ and $\vec{\beta}$ and the equalities (6.1) we have $\beta_4(0) = -\alpha_4(0)$. Moreover

$$\beta_4(0) = 4\left(\int_0^\infty (\delta_2 y_2 + \delta_4 y_5)\phi^2 y_2\,dx - \int_0^\infty (\delta_1 y_1 + \delta_3 y_3)\phi^2 y_2\,dx\right) = 4(I_1 - I_2). \tag{6.4}$$

We evaluate I_1 as

$$I_1 = 4\Bigg(\int_0^\infty \phi^2(x)\,y_2^2(x)\,dx \int_x^\infty \phi^2(s)\,y_2(s)\,y_5(s)\,ds \\ - \int_0^\infty \phi^2(x)\,y_2(x)\,y_5(x)\,dx \int_x^\infty \phi^2(s)\,y_2^2(s)\,ds\Bigg) = \frac{1}{27}(4-3\ln 2).$$

The integral I_2 takes the form

$$I_2 = \frac{4}{W_0}\Bigg(\int_0^\infty \phi^2(x)\,y_1(x)\,y_2(x)\,dx \int_x^\infty \phi^2(s)\,y_2(s)\,y_3(s)\,ds \\ - \int_0^\infty \phi^2(x)\,y_2(x)\,y_3(x)\,dx \int_x^\infty \phi^2(s)\,y_1(s)\,y_2(s)\,ds\Bigg) = \frac{\Im(\mathcal{G})}{4\alpha\,(1-\rho^2)},$$

where $\Im(\mathcal{G})$ is the imaginary part of the integral $\mathcal{G}$ defined by

$$\mathcal{G} = \int_0^\infty \operatorname{sech}^3(x)\big(i\alpha + \tanh x\big)e^{-i\alpha x}dx \int_x^\infty \operatorname{sech}^3(s)\big(i\alpha - \tanh s\big)e^{i\alpha s}ds.$$

Substituting these values into (6.2) yields

$$\begin{aligned} E_{BB}(\lambda_+,0) = \quad & E_\lambda(\lambda_+,0)\Bigg[\left(\tfrac{8(2-3\rho^2)}{9\alpha\,(1-\rho^2)} - \tfrac{2\alpha(1-\rho^2)\pi^2}{9}\operatorname{sech}^2\tfrac{\alpha\pi}{2}\right) \\ & + i\left(\tfrac{44}{27} - 2\tfrac{\Im(\mathcal{G})}{\alpha\,(1-\rho^2)} + \tfrac{2\pi\operatorname{sech}\frac{\alpha\pi}{2}}{9}\big(1+2\alpha\textstyle\int_0^\infty \operatorname{sech}^3(x)\,\sin(\alpha\,x)\,dx\big)\right)\Bigg]. \end{aligned} \tag{6.5}$$

Taking the real part of $\frac{E_{BB}}{E_\lambda}$, we arrive at the aforementioned result (4.15).

7. Discussion. We have described in detail the sequence of bifurcations characterizing the polarizational mode instability in the limit of weak birefringence. These occur in several families of NLS systems where nonlinear coupling occurs not only through amplitude via cross phase modulation but also through complex phase via the FWM. We find parameter regimes in which both the fast and the slow waves are linearly stable. The instability of the fast wave manifests itself as a rotation of polarization angle, the parameter α in (2.2), leading to a subsequent chaotic wobble. The impact of this instability on soliton collision, or more generally the modulational stability of the two soliton, is a subject of further study.

The Dirichlet expansion and the Evans function have shown themselves to be an efficient pair of analytical tools. The sharp resolution of the fourth root manifested in the eigenpath pinching of Figure 1(b) at the origin required many accurate evaluations of the Evans function. Moreover, the detailed resolution of the stability diagram (Figure 3) would have been impractical by more direct methods. Although the integrability of the unperturbed Manakov equation was reflected in the closed form expression of the Dirichlet expansion in that case, the integrability was not directly exploited, and the Dirichlet expansion may well find useful application to Evans function calculations for nonintegrable problems where significant information is available about the stable and unstable manifolds of the traveling wave forms.

Acknowledgments. We are thankful to T. Kapitula and B. Sandstede for many helpful discussions.

REFERENCES

[1] N. Akhmediev and A. Ankiewicz, *Solitons: Nonlinear Pulses and Beams*, Chapman and Hall, London, 1997.

[2] N. Akhmediev, A. Buryak, J.M. Soto-Crespo, and D. Anderson, *Phase-locked stationary soliton states in birefringent nonlinear optical fibers*, J. Opt. Soc. Amer. B, 12 (1995), pp. 434–439.

[3] N. Akhmediev, V. Eleonskii, E. Kulagin, and L. Shilnikov, *Steady state pulses in a birefringent nonlinear optical fiber: Soliton multiplication processes*, Sov. Tech. Phys. Lett, 15 (1989), pp. 587–588.

[4] J.C. Alexander, R.A. Gardner, and C.K.R.T. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[5] J.C. Alexander, M.G. Grillakis, C.K.R.T. Jones, and B. Sandstede, *Stability of pulses on optical fibers with phase-sensitive amplifiers*, Z. Angew. Math. Phys., 48 (1997), pp. 175–192.

[6] K.J. Blow, N.J. Doran, and D. Wood, *Polarization instabilities for solitons in birefringent fibers*, Opt. Lett., 12 (1987), pp. 202–204.

[7] D.N. Christodoulides and R.I. Joseph, *Vector solitons in birefringent nonlinear dispersive media*, Opt. Lett. 13 (1988), pp. 53–55.

[8] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[9] G. Evangelides, L. Mollenauer, J. Gordon, and N. Bergano, *Polarization multiplexing with solitons*, J. Lightwave Tech., 10 (1992), pp. 28–35.

[10] A. Hasegawa and F. Tappert, *Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers.* I. *Anomalous dispersion,* Appl. Phys. Lett., 23 (1973), pp. 142–144.

[11] R.A. Gardner and K. Zumbrun, *The gap lemma and geometric critera for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.

[12] M. Grillakis, *Analysis of the linearization around a critical point of an infinite dimensional Hamiltonian system*, Comm. Pure Appl. Math., 43 (1990), pp. 299–333.

[13] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry,* II, J. Funct. Anal., 94 (1990), pp. 308–348.

[14] C.R.T. Jones, *Instability of standing waves for nonlinear Schrödinger type equations*, Ergodic Theory Dynam. Systems, 8 (1988), pp. 119–138.

[15] C.R.T. Jones, *Stability of travelling wave solutions of the Fitzhugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[16] T. Kapitula, *The Evans function and generalized Melnikov integrals*, SIAM J. Math. Anal., 30 (1999), pp. 273–297.

[17] T. Kapitula and J. Rubin, *Existence and stability of standing hole solutions to complex Ginzburg-Landau equations*, Nonlinearity, 13 (2000), pp. 77–112.

[18] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, 124 (1998), pp. 58–103.

[19] D. Kaup, *Second-order perturbations for solitons in optical fibers*, Phys. Rev. A, 44 (1991), pp. 4582–4590.

[20] D. Kaup and T. Lakoba, *Perturbation theory for the Manakov soliton and its application to pulse propagation in randomly birefringent fibers*, Phys. Rev. E (3), 56 (1997), pp. 6147–6165.

[21] D. Kaup and B.A. Malomed, *Soliton trapping and daughter waves in the Manakov model*, Phys. Rev. A, 48 (1993), pp. 599–604.

[22] Y. Li and K. Promislow, *Structural stability of non-groundstate traveling waves of coupled nonlinear Schrödinger systems*, Phys. D, 124 (1998), pp. 137–165.

[23] S.V. Manakov, *On the theory of two-dimensional stationary self-focusing of electro-magnetic waves*, Soviet Physics JETP, 38 (1974), pp. 248–253.

[24] C. Menyuk, *Stability of solitons in birefringent optical fibers,* I. *Equal propagation amplitudes*, Opt. Lett., 12 (1987), pp. 614–616.

[25] C. Menyuk, *Pulse propagation in an elliptically birefringent optical Kerr medium*, IEEE J. Quant. Electron., 25 (1989), pp. 2674–2682.

[26] D. Mihalache, D. Mazilu, and L. Torner, *Stability of walking vector solitons*, Phys. Rev. Lett., 81 (1998), pp. 4353–4356.

[27] L.F. Mollenauer, J.P. Gordon, and F. Heismann, *Polarization scattering by soliton-soliton collisions*, Opt. Lett., 20 (1995), pp. 2060–2063.

[28] D.J. Muraki and W.L. Kath, *Hamiltonian dynamics of solitons in optical fibers*, Phys. D, 48 (1991), pp. 53–64.

[29] Y. Nogami and C.S. Warke, *Soliton solutions of multicomponent nonlinear Schrödinger equation*, Phys. Lett., 59 (1976), pp. 251–253.

[30] R.L. Pego and M.I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.

[31] J.M. Soto-Crespo, N.N. Akhmediev, and A. Ankiewicz, *Stationary soliton-like pulses in birefringent optical fibers*, Phys. Rev. E, 51 (1995), pp. 3547–3555.

[32] C. Sophocleous and D.F. Parker, *Pulse collisions and polarization conversion for optical fibers*, Opt. Comm., 112 (1994), p. 214.

[33] M.W. Tratnik and J.E. Sipe, *Bound solitary waves in birefringent optical fiber*, Phys. Rev. A, 38 (1988), pp. 2011–2017.

[34] T. Ueda and W. Kath, *Dynamics of optical pulses in randomly birefringent fibers*, Phys. D, 55 (1992), pp. 166–181.

[35] P.A. Wai and C.R. Menyuk, *Polarization mode dispersion, decorrelation, and diffusion in optical fibers with randomly varying birefringence*, J. Lightwave Tech., 14 (1996), pp. 148–157.

[36] M. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal, 16 (1985), pp. 472–491.

[37] M. Weinstein, *Liapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.

[38] J. Yang, *Multisoliton perturbation theory for the Manakov equations and its applications to nonlinear optics*, Phys. Rev. E (3), 59 (1999), pp. 2393–2405.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1374–1395

CONVERGENCE OF MEISSNER MINIMIZERS OF THE GINZBURG–LANDAU ENERGY OF SUPERCONDUCTIVITY AS $\kappa \to +\infty$[*]

A. BONNET[†], S. J. CHAPMAN[‡], AND R. MONNEAU[§]

Abstract. The Meissner solution of a smooth cylindrical superconducting domain subject to a uniform applied axial magnetic field is examined. Under an additional convexity condition the uniqueness of the Meissner solution is proved. It is then shown that it is a local minimizer of the Ginzburg–Landau energy $\mathcal{E}_\kappa$, For applied fields less than a critical value, the existence of the Meissner solution is proved for large enough Ginzburg–Landau parameter κ. Moreover it is proved that the Meissner solution converges to a local minimizer of a certain energy $\mathcal{E}_\infty$ in the limit as $\kappa \to \infty$. Finally, it is proved that for κ large enough the Meissner solution is not a global minimizer of $\mathcal{E}_\kappa$.

Key words. nonlinear elliptic PDE, type II superconductors, local and global minimizers, asymptotic behavior, inverse function theorem

AMS subject classifications. 35B25, 35B30, 35B45, 35E10, 35J55, 35Q40, 82D55

PII. S0036141098346165

1. Introduction.

At temperatures below their critical temperature, type II superconductors can exist in one of three different states, depending on the strength of the applied magnetic field. For low magnetic fields they are in what is known as the Meissner state, in which the magnetic field is excluded from the interior except in thin boundary layers whose thickness is known as the penetration depth, denoted by λ. At intermediate applied magnetic fields the magnetic field penetrates the sample in the form of quantized flux tubes, usually known as vortices, since they are each circled by a vortex of superconducting current. The cores of these tubes comprise nonsuperconducting (normal) material, and are of a radius ξ, which is known as the coherence length. An important physical parameter is the ratio of these two lengthscales, $\kappa = \lambda/\xi$, which is known as the Ginzburg–Landau parameter. For type II superconductors $\kappa > 1/\sqrt{2}$, but κ is typically much larger than this, especially in superconducting alloys. We note that all the high-temperature superconducting materials discovered to date have very large (typically 50 to 100) values of κ.

As the applied magnetic field increases further, the density of vortices increases. For large applied magnetic fields the material becomes completely normal.

The transition from the vortex state to the normal state occurs at the upper critical field, H_{c_2}, and is second order and therefore reversible. On the other hand, the transition from the Meissner state to the vortex state is first order and exhibits hysteresis. The magnetic field at which the vortex state becomes theoretically energetically favorable is known as the lower critical field, H_{c_1}, but there is both "superheating" and "supercooling" around this field. The superheating of the Meissner state is the subject of this paper.

[*] Received by the editors October 26, 1998; accepted for publication (in revised form) September 23, 1999; published electronically May 19, 2000. This work was initiated by a grant from the British Council.

http://www.siam.org/journals/sima/31-6/34616.html

[†] Université de Cergy-Pontoise, 95011 Cergy-Pointoise cedex, France.

[‡] Mathematical Institute, Oxford University, 24-29 St. Giles', Oxford OX1 3LB, UK (chapmann@maths.ox.ac.uk). Supported by a Royal Society University Research Fellowship.

[§] Ecole des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, Cite Descartes, 77455 Marne la Vallee, France.

We study the Meissner state for finite κ and in the limit as $\kappa \to \infty$, in which case the model is substantially simplified. The layout of the paper is as follows. In the remainder of the introduction we introduce the Ginzburg–Landau theory of superconductivity and state our main results. We also establish the convergence of the Meissner solution to the solution of the simplified problem as $\kappa \to \infty$.

In section 3 we prove the uniqueness of the Meissner solution, while in section 4 we establish its existence for magnetic fields less than a critical value and for large enough κ. In section 5 we consider global and local minimizers of the Ginzburg–Landau (GL) free energy. We show that the Meissner solution is a local minimizer for magnetic fields below a (second) critical value, but that for large κ and nonzero applied magnetic field it is not the global minimizer. Finally, we present our conclusions.

2. The Ginzburg–Landau model.

We consider a superconductor material occupying a domain $\Sigma \subset \mathbb{R}^3$ in a uniform exterior magnetic field $H_0\boldsymbol{e_3}$. The state of the superconductor is characterized by a complex order parameter Ψ (defined on Σ) such that $|\Psi|^2$ represents the number density of superconducting electrons, which may be thought of as a kind of "macroscopic wavefunction," and the magnetic vector potential $\boldsymbol{A}$, which is such that the magnetic field is given by

$$\boldsymbol{H} = \mathrm{curl}\ \boldsymbol{A}.$$

In the theory introduced by Ginzburg and Landau [9] the equilibrium state of the superconductor is given by the minimizer of the Ginzburg–Landau energy (for a review of the theory the reader may consult [4, 6, 7]):

$$\mathcal{E} = \int_\Sigma \left|\left(\frac{1}{\kappa}\nabla - i\boldsymbol{A}\right)\Psi\right|^2 + \frac{(|\Psi|^2-1)^2}{2} + |\,\mathbf{curl}\ \boldsymbol{A} - H_0\boldsymbol{e_3}|^2\,dV + \int_{\mathbb{R}^3\setminus\Sigma} |\,\mathbf{curl}\ \boldsymbol{A} - H_0\boldsymbol{e_3}|^2\,dV. \tag{2.1}$$

Here κ is the Ginzburg–Landau parameter of the introduction. The energy (2.1) is in the usual nondimensional form in which $|\Psi| = 1$ represents wholly superconducting material, $|\Psi| = 0$ represents wholly normal material, and lengths have been scaled with the penetration depth λ, which is the typical lengthscale for variations in the magnetic potential. The vortex core radius is then given by $1/\kappa$.

The Euler–Lagrange equations corresponding to (2.1) are the celebrated Ginzburg–Landau equations

$$\left(\frac{1}{\kappa}\nabla - i\boldsymbol{A}\right)^2\Psi = \left(|\Psi|^2-1\right)\Psi \qquad \text{in } \Sigma, \tag{2.2}$$

$$(\mathrm{curl})^2\boldsymbol{A} = \frac{i}{2\kappa}\left(\Psi\nabla\Psi^* - \Psi^*\nabla\Psi\right) - |\Psi|^2\boldsymbol{A} \qquad \text{in } \Sigma, \tag{2.3}$$

$$(\mathrm{curl})^2\boldsymbol{A} = \boldsymbol{0} \qquad \text{in } \mathbb{R}^3\setminus\Sigma, \tag{2.4}$$

$$\boldsymbol{n}\cdot\left(\frac{1}{\kappa}\nabla - i\boldsymbol{A}\right)\Psi = 0 \qquad \text{on } \partial\Sigma, \tag{2.5}$$

$$[\boldsymbol{n}\wedge\boldsymbol{A}] = \boldsymbol{0}, \tag{2.6}$$

$$[\boldsymbol{n}\wedge\mathrm{curl}\ \boldsymbol{A}] = \boldsymbol{0}, \tag{2.7}$$

$$\mathrm{curl}\ \boldsymbol{A} \to H_0\boldsymbol{e_3} \qquad \text{as } |\boldsymbol{x}| \to \infty, \tag{2.8}$$

where [ ] represents the jump in the enclosed quantity across $\partial\Sigma$, and $\boldsymbol{n}$ is the unit outward normal to Σ. We consider the case where the superconductor has the shape of a long cylinder, mathematically idealized by $\Sigma = \Omega \times \mathbb{R}$ where the section of the cylinder $\Omega \subset \mathbb{R}^2$ is a bounded smooth open set, and the axis of the cylinder is parallel to $\boldsymbol{e_3}$. Then the energy (2.1) makes no sense, but assuming that the problem is invariant by translation in the direction $\boldsymbol{e_3}$ of the axis of the cylinder, we can simply redefine the energy as an integral over the cross-section $\mathbb{R}^2$. If we assume also that the magnetic field lies only in the $\boldsymbol{e}_3$ direction, then (2.4) and (2.8) imply that the magnetic field is constant and equal to $H_0\boldsymbol{e}_3$ in $\mathbb{R}^2\backslash\Omega$. Then the continuity conditions on $\partial\Omega$ can be replaced by $\boldsymbol{H} = H_0\boldsymbol{e}_3$ there.

Vortices correspond to zeros of the order parameter Ψ, while the Meissner solution is a solution in which $|\Psi| > 0$ on $\overline{\Omega}$. In this case we can make the change of gauge

$$\begin{cases} \Psi = fe^{i\chi}, f > 0 \\ \mathbf{Q} = \boldsymbol{A} - \frac{1}{\kappa}\nabla\chi, \end{cases}$$

where χ is uniquely defined (up to $2k\pi$, $k \in \mathbf{Z}$) if, for example, Ψ is smooth. Then we define the new energy for $\kappa \in (0, +\infty]$:

$$\mathcal{E}_\kappa(f, \mathbf{Q}) = \int_\Omega \frac{|\nabla f|^2}{\kappa^2} + |\text{curl } \mathbf{Q} - H_0|^2 + G(f, \mathbf{Q}),$$

where

$$G(f, \mathbf{Q}) = f^2|\mathbf{Q}|^2 + \frac{(f^2 - 1)^2}{2}$$

and curl $\mathbf{Q} = \partial_1 Q_2 - \partial_2 Q_1$ with $\mathbf{Q} = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$. For $\kappa = +\infty$, we take

$$\mathcal{E}_\infty(f, \mathbf{Q}) = \int_\Omega |\text{curl } \mathbf{Q} - H_0|^2 + G(f, \mathbf{Q}).$$

Let us remark that $\frac{1}{2}G''(f, \mathbf{Q}) \cdot (g, \boldsymbol{q})^2 = (f\boldsymbol{q} + 2g\mathbf{Q})^2 + 3g^2(f^2 - |\mathbf{Q}|^2 - \frac{1}{3})$, so that G is convex if $f^2 - |\mathbf{Q}|^2 \geq \frac{1}{3}$.

2.1. A classical difficulty and the mathematical framework. Let V be the space (to be defined) on which we want to study the energy $\mathcal{E}_\kappa$. Let us remark that the functional $\mathcal{E}_\kappa$ is convex on the convex set $K_0 = \{(f, \mathbf{Q}) \in V,\ f^2 - |\mathbf{Q}|^2 \geq \frac{1}{3}\}$.

We are interested in the solutions $(f_\kappa, \mathbf{Q}_\kappa)$ such that

$$\inf_{(f,\mathbf{Q})\in K_0} \mathcal{E}_\kappa(f, \mathbf{Q}) \tag{2.9}$$

is attained by $(f_\kappa, \mathbf{Q}_\kappa)$ and $(f_\kappa, \mathbf{Q}_\kappa)$ belongs to $\text{Int}(K_0)$, the interior of K_0. For the value H_0 of the exterior magnetic field, we will denote by $u_\kappa^{H_0}$ such a minimizing point $(f_\kappa, \mathbf{Q}_\kappa)$.

Then at least two choices seem natural:

(a) $V = V_4$, where $V_4 = \{(f, \mathbf{Q}) \in L^4 \times L^4,\ \nabla f \in L^2,\ \text{curl } \mathbf{Q} \in L^2\}$, with the norm $|(f, \mathbf{Q})|_{V_4} = |f|_{L^4} + |\mathbf{Q}|_{L^4} + |\nabla f|_{L^2} + |\text{curl } \mathbf{Q}|_{L^2}$, the space on which the energy $\mathcal{E}_\kappa$ is naturally defined.

(b) The constraint which defines K_0 requires an L^∞ control on f and $\mathbf{Q}$. Then it is natural to define $V = V_\infty$ where $V_\infty = \{(f, \mathbf{Q}) \in L^\infty \times L^\infty,\ \nabla f \in L^2,\ \text{curl } \mathbf{Q} \in L^2\}$, with the norm $|(f, \mathbf{Q})|_{V_\infty} = |f|_{L^\infty} + |\mathbf{Q}|_{L^\infty} + |\nabla f|_{L^2} + |\text{curl } \mathbf{Q}|_{L^2}$.

Let us remark that in each case, (a) or (b), the energy $\mathcal{E}_\kappa$ is defined and continuous on V. Generally the existence of a minimizer $u_\kappa^{H_0}$ is a consequence of the fact that

$\mathcal{E}_\kappa(f,\mathbf{Q}) \to +\infty$ as $|(f,\mathbf{Q})|_V \to +\infty$; but here in both cases, (a) and (b), this property is not clear. A second difficulty is that if $u_\kappa^{H_0}$ is a minimizer on K_0, how can we know that $u_\kappa^{H_0} \notin \partial K_0$ (and then $u_\kappa^{H_0} \in \mathrm{Int}(K_0)$)? A great difficulty in the case (a) is that if $u_\kappa^{H_0} \in \mathrm{Int}(K_0)$, then for a perturbation of applied fields h in a neighborhood of H_0, it is not possible to claim that u_κ^h stays in $\mathrm{Int}(K_0)$; on the contrary this is possible in the case (b). This difficulty is associated with the fact that K_0 is not closed when $V = V_4$, but is closed if $V = V_\infty$.

Then we chose $V = V_\infty$ and we define the convex closed sets

$$K_{\delta_0} = \Big\{(f,\mathbf{Q}) \in V,\ f^2 - |\mathbf{Q}|^2 \geq \frac{1}{3} + \delta_0\Big\},\ \delta_0 \geq 0.$$

Then $\mathrm{Int}(K_0) = \cup_{\delta_0>0} K_{\delta_0}$. In particular $\mathcal{E}_\kappa$ is strictly convex on K_{δ_0}. Moreover there exists $c = c(\delta_0) > 0$ such that if $(f,\mathbf{Q})$ is a critical point of $\mathcal{E}_\kappa$ and $(f,\mathbf{Q}), (f+\overline{f}, \mathbf{Q}+\overline{\mathbf{Q}}) \in K_{\delta_0}$, then

$$\mathcal{E}_\kappa(f+\overline{f}, \mathbf{Q}+\overline{\mathbf{Q}}) \geq \mathcal{E}_\kappa(f,\mathbf{Q}) + \int_\Omega \frac{|\nabla \overline{f}|^2}{\kappa^2} + |\mathrm{curl}\ \overline{\mathbf{Q}}|^2 + c(|\overline{f}|^2 + |\overline{\mathbf{Q}}|^2).$$

2.2. Main results.

Definition 2.1. *We say that $(f,\mathbf{Q}) \in V$ is a local minimizer of $\mathcal{E}_\kappa$ on V if $\exists \epsilon = \epsilon(f,\mathbf{Q}) > 0$ such that if $|(\overline{f},\overline{\mathbf{Q}})|_V < \epsilon$, then $\mathcal{E}_\kappa(f+\overline{f}, \mathbf{Q}+\overline{\mathbf{Q}}) \geq \mathcal{E}_\kappa(f,\mathbf{Q})$. The function $(f,\mathbf{Q})$ is said to be a strict local minimizer if, moreover, $\mathcal{E}_\kappa(f+\overline{f}, \mathbf{Q}+\overline{\mathbf{Q}}) > \mathcal{E}_\kappa(f,\mathbf{Q})$ for $(\overline{f},\overline{\mathbf{Q}}) \neq 0$. Finally, $(f,\mathbf{Q})$ is a global minimizer if we can take $\epsilon = +\infty$ in the definition of a local minimizer.*

An element $(f,\mathbf{Q}) \in V$ is said to be a critical point of $\mathcal{E}_\kappa$ on V if it satisfies the Euler–Lagrange equations

$$\left\{ \begin{array}{l} \frac{1}{\kappa^2}\Delta f = \frac{1}{2}G'_f \\ -\ \mathbf{curl}\ (\mathrm{curl}\ \mathbf{Q}) = \frac{1}{2}G'_{\mathbf{Q}} \end{array} \right| \text{ on } \Omega \tag{2.10}$$

with the boundary conditions (that we obtain formally)

$$\left\{ \begin{array}{l} \frac{\partial_n f}{\kappa^2} = 0 \\ \mathrm{curl}\ \mathbf{Q} = H_0 \end{array} \right| \text{ on } \partial\Omega. \tag{2.11}$$

We assume that $\partial\Omega \in C^{2,\alpha}$ for some $\alpha \in (0,1)$. Let us remark that we can deduce (2.11) rigorously as weak boundary conditions (see (3.2)). Then we have (see [13, 14, 15] for similar uniqueness results for the Meissner solution).

Theorem 2.2. *If $(f,\mathbf{Q})$ is a critical point of $\mathcal{E}_\kappa$ on V (a weak solution of* (2.10)–(2.11)*), and $(f,\mathbf{Q}) \in \mathrm{Int}(K_0)$, then $(f,\mathbf{Q})$ is unique. Moreover $(f,\mathbf{Q})$ is a strict local minimizer of $\mathcal{E}_\kappa$.*

In what follows we denote by $(f_\kappa, \mathbf{Q}_\kappa)$ this local minimizer when it exists. We will use the more precise and compact notation $u_\kappa^{H_0}$ to denote this solution under an exterior magnetic field H_0.

Let us recall that in [1], Berestycki, Bonnet, and Chapman have proved with help of an inverse function theorem that there exists a critical magnetic field H_0^* which only depends on Ω such that $\forall H_0 \in [0, H_0^*)$, $u_\infty^{H_0}$ exists; $u_\infty^{H_0^*} \in \partial K_0$; and $\forall H_0 > H_0^*$, $u_\infty^{H_0}$ does not exist. Then we have the following theorem.

Theorem 2.3. *There exists a positive continuous and nondecreasing function γ defined on $(0,+\infty]$ such that $\gamma(+\infty) = H_0^*$ and such that $\forall(\kappa, H_0) \in \mathcal{P} := \{(\kappa, H_0),\ \kappa \in$*

$(0,+\infty],\ 0 \leq H_0 < \gamma(\kappa)\}$, $u_\kappa^{H_0} \in \mathrm{Int}(K_0)$ exists. Moreover $(\kappa, H_0) \to u_\kappa^{H_0} \in V$ is continuous on $\mathcal{P}$. In particular this implies that for $0 \leq H_0 < H_0^*$: $u_\kappa^{H_0} \to u_\infty^{H_0}$ in V as $\kappa \to +\infty$.

Now we give a result which proves that for κ large enough, $u_\kappa^{H_0}$ is only a local minimizer and not a global minimizer.

Theorem 2.4. *$\forall H_0 \geq 0, \exists C = C(H_0) > 0$ such that*

$$\inf_{(f,\mathbf{Q})\in V} \mathcal{E}_\kappa(f,\mathbf{Q}) \leq \frac{C}{\sqrt{\kappa}}.$$

In particular for $\kappa = +\infty$, $\mathcal{E}_\infty$ has no global minimizer if $H_0 > 0$.

2.3. Proof of Theorem 2.3. We give here the proof of Theorem 2.3, because it is very short and shows the main idea of this article. The proof of Theorem 2.3 uses the following two propositions.

Proposition 2.5 (case $\kappa \in (0,+\infty)$). *For $\kappa_0 \in (0,+\infty)$ and $H_0 \geq 0$, if $u_{\kappa_0}^{H_0} \in \mathrm{Int}(K_0)$, then $\exists \epsilon = \epsilon(\kappa_0, H_0) > 0$, such that if $|h - H_0|, |\kappa - \kappa_0| < \epsilon$, then u_κ^h exists and $u_\kappa^h \in \mathrm{Int}(K_0)$. Moreover the map $(\kappa, h) \longmapsto u_\kappa^h \in V$ is continuous.*

We denote curl $\mathbf{Q}_\infty$ by H_∞ where $(f_\infty, \mathbf{Q}_\infty)$ is the solution $u_\infty^{H_0}$. In [10] it is proved that the map $H_0 \longmapsto \sup_{\overline{\Omega}} |\nabla H_\infty|$ is nondecreasing. Then $\forall \delta_0 > 0$, there exists a unique $H_{\delta_0}^* \in (0, H_0^*)$ such that $\forall h \in [0, H_{\delta_0}^*]$, $u_\infty^h \in K_{\delta_0}$, and $\forall h > H_{\delta_0}^*$, $u_\infty^h \notin K_{\delta_0}$.

We need to have an a priori L^∞-control on f and $\mathbf{Q}$, so we define for $M > 1$ the set

$$K_{\delta_0}^M = \left\{ (f,\mathbf{Q}) \in K_{\delta_0},\ f, |\mathbf{Q}| \leq M \text{ and } f \geq \frac{1}{M} \right\}.$$

Then we have the following proposition.

Proposition 2.6 (case κ close to $+\infty$). *For $\delta_0 > 0$, $\exists \kappa_0 = \kappa_0(\delta_0) > 0$ such that $\forall M > 1\ \forall \theta \in (0,1)$, $\exists C_\theta = C_\theta(\delta_0, M) > 0$ such that $\forall H_0 \in [0, H_{\delta_0}^*]\ \forall \kappa \geq \kappa_0$, then if $u_\kappa^{H_0}$ exists and $u_\kappa^{H_0} \in K_{\delta_0}^M$, then*

$$|u_\kappa^{H_0} - u_\infty^{H_0}|_V \leq \frac{C_\theta(\delta_0, M)}{\kappa^{\frac{1}{2}-\theta}}.$$

Given these two propositions, the proof of Theorem 2.3 can be made in three steps:

Step 1. Fix some $\theta \in (0, \frac{1}{2})$. Let $\delta_0 > 0$ be given. Choose $M_0 = M_0(\delta_0) > 1$ such that $\forall H_0 \in [0, H_{\delta_0}^*]$, $u_\infty^{H_0} \in K_{\delta_0}^{M_0}$. Then choose $\kappa_1 = \kappa_1(\delta_0) > 0$, large enough such that $\forall H_0 \in [0, H_{\delta_0}^*]$, then

$$\forall u \in V,\ \left(|u - u_\infty^{H_0}|_V \leq \frac{C_\theta(\frac{\delta_0}{4}, 4M_0)}{\kappa_1^{\frac{1}{2}-\theta}} \right) \Longrightarrow \left(u \in K_{\frac{\delta_0}{2}}^{2M_0} \right). \tag{2.12}$$

Step 2. Let $\kappa \geq \kappa_1$. We have $u_\kappa^0 = (1,0) \in \mathrm{Int}(K_0)$ and Proposition 2.5 allows us to build a u_κ^h for h in a neighborhood of 0, and with help of the control given by Proposition 2.6 and (2.12), u_κ^h stays in $K_{\frac{\delta_0}{2}}^{2M_0}$, and then we have existence of u_κ^h until $h = H_{\delta_0}^*$.

Step 3. For $H_0 \in [0, H_0^*)$, from Proposition 2.5, the map $\kappa \longmapsto u_\kappa^{H_0} \in V$ is continuous, and from Proposition 2.6, $u_\kappa^{H_0}$ admits the limit $u_\infty^{H_0}$ as $\kappa \to +\infty$. Consequently we can always choose a positive continuous and nondecreasing function γ as described in Theorem 2.3.

Theorem 2.3 is then proved. □

3. Uniqueness of Meissner solutions in an infinite cylinder under a convexity condition. Here we are interested in the uniqueness of Meissner solutions of the Ginzburg–Landau equations in infinite cylinder $\Sigma_\infty = \Omega \times \mathbb{R}$. The problem is invariant by translation, and we usually restrict the study to solutions which have this symmetry (which is the case in Theorem 2.2):

$$f = f(x_1, x_2), \qquad \mathbf{Q} = \begin{pmatrix} Q_1(x_1, x_2) \\ Q_2(x_1, x_2) \\ 0 \end{pmatrix}, \qquad (x_1, x_2) \in \partial\Omega.$$

Here we relax these conditions somewhat but not completely. We still suppose that the magnetic field outside the superconductor lies in the $\boldsymbol{e}_3$ direction, so that it is constant and equal to $H_0\boldsymbol{e}_3$, and the energy density in $\mathbb{R}^3\backslash\Sigma_\infty$ vanishes. However, we relax the symmetry assumption inside the superconductor, so that we prove the uniqueness (under a certain convexity condition) of solutions $(f, \mathbf{Q})$ defined in all the cylinder Σ_∞ of the form

$$f = f(x_1, x_2, x_3), \qquad \mathbf{Q} = \begin{pmatrix} Q_1(x_1, x_2, x_3) \\ Q_2(x_1, x_2, x_3) \\ Q_3(x_1, x_2, x_3) \end{pmatrix};$$

Theorem 2.2 will be a particular case of Theorem 3.1. As we have already mentioned, we cannot derive the Ginzburg–Landau equations on an infinite cylinder from an energy. Let us then take as a model the case of a periodic cylinder and derive rigorously the Euler–Lagrange equations. In this way we will find the minimal regularity assumption to impose on our solution to the infinite cylinder.

Then for $\Lambda > 0$, let the periodic cylinder $\Sigma_\Lambda = \Omega \times \Lambda\mathbf{S}^1$ where $\mathbf{S}^1 = \mathbf{Z}/(2\pi\mathbf{Z})$, and let f, $\mathbf{Q}$ be functions defined on $\overline{\Sigma}_\Lambda$. Then for $\kappa \in (0, +\infty]$, the Ginzburg–Landau energy is

$$\mathcal{E}_\kappa^\Lambda(f, \mathbf{Q}) = \int_{\Sigma_\Lambda} \frac{|\nabla f|^2}{\kappa^2} + |\ \mathbf{curl}\ \mathbf{Q} - H_0\boldsymbol{e}_\mathbf{3}|^2 + G(f, \mathbf{Q}).$$

Let us recall that $V = \{(f, \mathbf{Q}) \in L^\infty \times L^\infty,\ \nabla f \in L^2,\ \ \mathbf{curl}\ \mathbf{Q} \in L^2\}$. Then $\mathcal{E}_\kappa$ is defined on V. In particular, if $(f, \mathbf{Q})$ is an extremum of $\mathcal{E}_\kappa^\Lambda$, then we get the Euler–Lagrange equations in the distributional sense inside Σ_λ:

$$\left\{ \begin{array}{l} \frac{1}{\kappa^2}\Delta f = \frac{1}{2}G'_f \in L^\infty \\ -(\ \mathbf{curl}\ )^2\mathbf{Q} = \frac{1}{2}G'_\mathbf{Q} \in L^\infty \end{array} \right| \text{ in } \mathcal{D}'(\Sigma_\Lambda) \tag{3.1}$$

and the boundary conditions in $H^{-\frac{1}{2}}(\partial\Sigma_\Lambda)$:

$$\left\{ \begin{array}{l} \partial_n f = 0 \\ (\ \mathbf{curl}\ \mathbf{Q} - H_0\boldsymbol{e}_\mathbf{3}) \wedge n = 0. \end{array} \right. \tag{3.2}$$

The only difficulty is to justify the weak sense of the boundary conditions. It is possible to define these conditions by duality (for example $\langle \partial_n f, \phi \rangle_{H^{-\frac{1}{2}}(\partial\Sigma_\Lambda) \times H^{\frac{1}{2}}(\partial\Sigma_\Lambda)} = \int_{\Sigma_\Lambda} \Delta f \phi + \nabla f \cdot \nabla\phi$ for $\phi \in H^1(\Sigma_\Lambda)$). In particular see [8] for the definition of the tangential trace of a vector field $\boldsymbol{A} \in L^2$ with $\mathbf{curl}\ \boldsymbol{A} \in L^2$.

Minimal regularity assumption. In the following we assume at least the minimal regularity $f, \mathbf{Q} \in L^\infty(\Sigma_\infty)$, ∇f, $\mathbf{curl}\ \mathbf{Q} \in L^2_{loc}(\Sigma_\infty)$ (and $\mathbf{curl}\ ^2\mathbf{Q} \in L^2_{loc}(\Sigma_\infty)$ which is a consequence of the Ginzburg–Landau equations).

Then we have in the case of the infinite cylinder the following theorem.

Theorem 3.1. *For $\kappa \in (0, +\infty]$, if $(f, \mathbf{Q})$ satisfies* (3.1)–(3.2) *on $\Sigma_\infty = \Omega \times \mathbb{R}$, with the minimal regularity assumption and $(f, \mathbf{Q}) \in K_{\delta_0} = \{(f, \mathbf{Q}) \in V,\ f^2 - |\mathbf{Q}|^2 \geq \frac{1}{3} + \delta_0\}, \delta_0 > 0$, then $(f, \mathbf{Q})$ is unique and*

$$\begin{cases} Q_3 \equiv 0, \\ f \textit{ and } \mathbf{Q} \textit{ are independent on } x_3. \end{cases}$$

In particular $\boldsymbol{H} = H(x_1, x_2)\boldsymbol{e_3}$ where $\boldsymbol{H} = \mathbf{curl}\ \mathbf{Q}$, $(x_1, x_2) \in \Omega$.

Now we can identify the solutions $(f, \mathbf{Q})$ on Σ_∞ as functions defined on Ω, and we have the following proposition.

Proposition 3.2. *The solutions $(f, \mathbf{Q})$ of Theorem* 3.1 *are analytic in Ω. On $\overline{\Omega}$, the solutions $(f, \mathbf{Q})$ have at least the regularity $C^{1+m,\alpha}$ if $\partial\Omega \in C^{2+m,\alpha}$, $m \geq 0$.*

Proof of Proposition 3.2. If $f, \mathbf{Q} \in L^\infty(\Omega)$, ∇f, $\mathbf{curl}\ \mathbf{Q} \in L^2(\Omega)$, then for finite κ

$$\begin{cases} \frac{1}{\kappa^2}\Delta f = \frac{1}{2}G'_f \in L^\infty \text{ in } \mathcal{D}'(\Omega), \\ \partial_n f = 0 \text{ in } H^{-\frac{1}{2}}(\partial\Omega). \end{cases}$$

Then (see [2] for the regularity of weak solutions) $f \in H^2(\Omega)$, and by the standard L^p-elliptic theory, $f \in W^{2,p}$ and therefore $f \in C^{1,\theta}$ (Schauder theory). Let us remark that the equation on $\mathbf{Q}$ is elliptic in dimension 1 but is not elliptic in dimension $n \geq 2$. So it is interesting to obtain an elliptic formulation, and it was done in [1] introducing the magnetic field $H = \text{curl}\ \mathbf{Q} = \partial_1 Q_2 - \partial_2 Q_1$. Using the explicit form of G it is easy to obtain an elliptic equation satisfied by H. Writing $-\ \mathbf{curl}$ (curl $\mathbf{Q}$) $= \frac{1}{2}G'_{\mathbf{Q}}$ as $-\ \mathbf{curl}\ H = f^2\mathbf{Q}$ where $\mathbf{curl}\ H = \left(\begin{smallmatrix} \partial_2 H \\ -\partial_1 H \end{smallmatrix}\right)$, we see that $H = \text{curl}\ \mathbf{Q} = \text{curl}\ (-\frac{\mathbf{curl}\ H}{f^2}) = \nabla \cdot (\frac{\nabla H}{f^2})$. Thus

$$\begin{cases} \nabla \cdot (\frac{\nabla H}{f^2}) - H = 0 \text{ in } \mathcal{D}'(\Omega), \\ H = H_0 \text{ in } H^{-\frac{1}{2}}(\partial\Omega); \end{cases} \tag{3.3}$$

$\mathbf{Q}$ may be found by the inverse formula $\mathbf{Q} = -\frac{\mathbf{curl}\ H}{f^2}$. Then $H \in C^{2,\theta}$ (see [2]). Consider the elliptic system

$$\begin{cases} \frac{1}{\kappa^2}\Delta f = \frac{1}{2}G'_f(f, -\frac{\mathbf{curl}\ H}{f^2}), \\ \partial_n f = 0, \\ \nabla(\frac{\nabla H}{f^2}) - H = 0, \\ H = H_0. \end{cases} \tag{3.4}$$

A classical bootstrap argument permits us to see that if $(f, H) \in C^{m,\alpha} \times C^{m+1,\alpha}$ with $m \geq 1$, then $(f, H) \in C^{m+1,\alpha} \times C^{m+2,\alpha}$, and then $f, H \in C^\infty(\Omega)$. Moreover because the elliptic (nonlinear) system is analytic, the solutions (f, H) are analytic on Ω (result of Morrey; see [11]). In particular if $\partial\Omega$ is analytic, then (f, H) is analytic on $\overline{\Omega}$. Finally we find the regularity on $\mathbf{Q}$, taking $\mathbf{Q} = -\frac{\mathbf{curl}\ H}{f^2}$. For infinite κ, the proof can be found in [1]. This ends the proof.

Proof of Theorem 3.1. Let $(f_j, \mathbf{Q}_j)$, $j = 1, 2$, be two solutions in K_{δ_0} of (3.1)–(3.2) on $\Sigma_\infty = \Omega \times \mathbb{R}$. Let $\psi(x_3)$ be a function with compact support. We multiply the equation $\frac{1}{\kappa^2}\Delta(f_2-f_1)=\frac{1}{2}[G'_f]^{f_2,\mathbf{Q}_2}_{f_1,\mathbf{Q}_1}$ by $\psi(f_2-f_1)$, and the equation $-(\textbf{ curl })^2(\mathbf{Q}_2-\mathbf{Q}_1)=\frac{1}{2}[G'_{\mathbf{Q}}]^{f_2,\mathbf{Q}_2}_{f_1,\mathbf{Q}_1}$ by $\psi(\mathbf{Q}_2-\mathbf{Q}_1)$. Then by integration on $\Sigma=\Sigma_\infty$ we obtain

$$\int_\Sigma\left(\frac{1}{\kappa^2}\nabla(\psi(f_2-f_1))\cdot\nabla(f_2-f_1)+\textbf{ curl }(\psi(\mathbf{Q}_2-\mathbf{Q}_1))\cdot\textbf{ curl }(\mathbf{Q}_2-\mathbf{Q}_1)\right.$$
$$\left.+\frac{1}{2}[G']^{f_2,\mathbf{Q}_2}_{f_1,\mathbf{Q}_1}\cdot\psi\begin{pmatrix} f_2-f_1\\ \mathbf{Q}_2-\mathbf{Q}_1\end{pmatrix}\right)=0.$$

Here we have integrated by parts one time in f, and one time in $\mathbf{Q}$, using the equality $\int_\Sigma(\textbf{ curl }\boldsymbol{A})\cdot\boldsymbol{B}=\int_\Sigma\boldsymbol{A}\cdot(\textbf{ curl }\boldsymbol{B})+\int_{\partial\Sigma}(\boldsymbol{A}\wedge\boldsymbol{B})\cdot\boldsymbol{n}$. Moreover the boundary terms obtained for f and for $\mathbf{Q}$ are zero because of (3.2). Then we obtain (using $\textbf{curl }(f\boldsymbol{A})=\nabla f\wedge\boldsymbol{A}+f\textbf{ curl }\boldsymbol{A}$)

$$\int_\Sigma\psi\left\{\frac{1}{\kappa^2}|\nabla(f_2-f_1)|^2+|\textbf{ curl }(\mathbf{Q}_2-\mathbf{Q}_1)|^2+\frac{1}{2}[G']^{f_2,\mathbf{Q}_2}_{f_1,\mathbf{Q}_1}\cdot\begin{pmatrix} f_2-f_1\\ \mathbf{Q}_2-\mathbf{Q}_1\end{pmatrix}\right\}$$
$$=\int_\Sigma\left(-\frac{f_2-f_1}{\kappa}\right)\cdot\left(\frac{\nabla\psi}{\sqrt{\psi}}\right)\cdot\left(\sqrt{\psi}\frac{\nabla(f_2-f_1)}{\kappa}\right)$$
$$-\left(\frac{\nabla\psi}{\sqrt{\psi}}\wedge(\mathbf{Q}_2-\mathbf{Q}_1)\right)\cdot\left(\sqrt{\psi}\textbf{ curl }(\mathbf{Q}_2-\mathbf{Q}_1)\right).\tag{3.5}$$

But G is strictly convex on K_{δ_0}, and from the inequality of convexity we deduce that

$$\exists c=c(\delta_0)>0\text{ such that }\frac{1}{2}[G']^{f_2,\mathbf{Q}_2}_{f_1,\mathbf{Q}_1}\cdot\begin{pmatrix} f_2-f_1\\ \mathbf{Q}_2-\mathbf{Q}_1\end{pmatrix}\geq c((f_2-f_1)^2+(\mathbf{Q}_2-\mathbf{Q}_1)^2).$$

On the other hand we have the L^∞-estimates on the right-hand side of (3.5): $|f_2-f_1|\leq C$, $|\mathbf{Q}_2-\mathbf{Q}_1|\leq C$. Then from the Cauchy–Schwarz inequality we have

$$\int_\Sigma\psi\left\{\frac{|\nabla(f_2-f_1)|^2}{\kappa^2}+|\textbf{ curl }(\mathbf{Q}_2-\mathbf{Q}_1)|^2+\delta((f_2-f_1)^2+(\mathbf{Q}_2-\mathbf{Q}_1)^2)\right\}$$
$$\leq C\left(\int_\Sigma\frac{|\nabla\psi|^2}{\psi}\right)^{\frac{1}{2}}\left\{\left(\int_\Sigma\psi\frac{|\nabla(f_2-f_1)|^2}{\kappa^2}\right)^{\frac{1}{2}}+\left(\int_\Sigma\psi|\textbf{ curl }(\mathbf{Q}_2-\mathbf{Q}_1)|^2\right)^{\frac{1}{2}}\right\}.$$

But if we take $\psi(x)=\psi_0(\lambda x_3)$, where $\psi_0\geq 0$ has a support in $[-2,2]$, is equal to 1 on $[-1,1]$, then using the fact that $\int_\Sigma\frac{|\nabla\psi|^2}{\psi}=\lambda\int_\Sigma\frac{|\nabla\psi_0|^2}{\psi_0}$, we deduce as $\lambda\to 0$ that

$$\int_\Sigma\frac{|\nabla(f_1-f_2)|^2}{\kappa^2}+|\textbf{ curl }(\mathbf{Q}_2-\mathbf{Q}_1)|^2+c((f_2-f_1)^2+(\mathbf{Q}_2-\mathbf{Q}_1)^2)=0.$$

Consequently $f_2=f_1$ and $\mathbf{Q}_2=\mathbf{Q}_1$.

Now if $(f,\mathbf{Q})$ is a solution, for $t\in\mathbb{R}$ let $f^t(x_1,x_2,x_3)=f(x_1,x_2,x_3+t)$, and $\mathbf{Q}^t(x_1,x_2,x_3)=\mathbf{Q}(x_1,x_2,x_3+t)$. It is clear that $(f^t,\mathbf{Q}^t)$ is a solution, and from the uniqueness result we deduce that f and $\mathbf{Q}$ do not depend on the coordinate x_3.

Now let $\tilde{\mathbf{Q}}=\begin{pmatrix} Q_1(x_1,x_2)\\ Q_2(x_1,x_2)\\ 0\end{pmatrix}$. Then (for every fixed Λ) we have by construction

$$\mathcal{E}^\Lambda_\kappa(f,\mathbf{Q}+\overline{\mathbf{Q}})\leq\mathcal{E}^\Lambda_\kappa(f,\mathbf{Q})\text{ where }\overline{\mathbf{Q}}=\tilde{\mathbf{Q}}-\mathbf{Q}.\tag{3.6}$$

We conclude with the help of the following lemma.

LEMMA 3.3. *Assume that* $(f,\mathbf{Q})\in V$ *is a solution of* (3.1)–(3.2) *on* Σ_Λ. *Consider* $(\overline{f},\overline{\mathbf{Q}})\in V$, *and assume that* $(f,\mathbf{Q})$ *and* $(f+\overline{f},\mathbf{Q}+\overline{\mathbf{Q}})$ *are in* K_{δ_0}. *Then we have*

$$\mathcal{E}_\kappa^\Lambda(f+\overline{f},\mathbf{Q}+\overline{\mathbf{Q}})\geq\mathcal{E}_\kappa^\Lambda(f,\mathbf{Q})+\int_{\Sigma_\Lambda}\frac{|\nabla\overline{f}|^2}{\kappa^2}+|\ \mathbf{curl}\ \overline{\mathbf{Q}}|^2+c(|\overline{f}|^2+|\overline{\mathbf{Q}}|^2). \tag{3.7}$$

From (3.6), this lemma implies (with $\overline{f}=0$) that $\int_{\Sigma_\Lambda}|\ \mathbf{curl}\ \overline{\mathbf{Q}}|^2+c|\overline{\mathbf{Q}}|^2=0$, so that $\overline{\mathbf{Q}}=0$. Hence $Q_3=0$. Then $\boldsymbol{H}=\ \mathbf{curl}\ \mathbf{Q}=H(x_1,x_2)\boldsymbol{e_3}$, and Theorem 3.1 is proved.

Proof of Lemma 3.3. We have

$$\begin{aligned}\mathcal{E}_\kappa(f+\overline{f},\mathbf{Q}+\overline{\mathbf{Q}})&=\mathcal{E}_\kappa(f,\mathbf{Q})+\int_{\Sigma_\Lambda}\left\{\frac{2}{\kappa^2}\nabla f\cdot\nabla\overline{f}+\frac{1}{\kappa^2}|\nabla\overline{f}|^2\right.\\&\left.+2(\ \mathbf{curl}\ \overline{\mathbf{Q}})\cdot(\ \mathbf{curl}\ \mathbf{Q}-H_0\boldsymbol{e_3})+|\ \mathbf{curl}\ \overline{\mathbf{Q}}|^2+G(f+\overline{f},\mathbf{Q}+\overline{\mathbf{Q}})-G(f,\mathbf{Q})\right\}.\end{aligned}$$

Moreover, from Taylor's formula with integral remainder, we obtain

$$G(f+\overline{f},\mathbf{Q}+\overline{\mathbf{Q}})-G(f,\mathbf{Q})=G'_{(f,\mathbf{Q})}(f,\mathbf{Q})\cdot(\overline{f},\overline{\mathbf{Q}})+\int_0^1(1-t)(\overline{f},\overline{\mathbf{Q}})\cdot G''(f,\mathbf{Q}+t\overline{\mathbf{Q}})\cdot(\overline{f},\overline{\mathbf{Q}})dt,$$

which gives, after an integration by parts,

$$\begin{aligned}\mathcal{E}_\kappa(f+\overline{f},\mathbf{Q}+\overline{\mathbf{Q}})&=\mathcal{E}_\kappa(f,\mathbf{Q})+\int_{\Sigma_\Lambda}\frac{1}{\kappa^2}|\nabla\overline{f}|^2+|\ \mathbf{curl}\ \overline{\mathbf{Q}}|^2\\&+\int_0^1dt(1-t)\left(\int_{\Sigma_\Lambda}(\overline{f},\overline{\mathbf{Q}})\cdot G''(f+t\overline{f},\mathbf{Q}+t\overline{\mathbf{Q}})\cdot(\overline{f},\overline{\mathbf{Q}})\right)\\&+2\left\{\int_{\Sigma_\Lambda}\overline{f}\cdot\left(-\frac{1}{\kappa^2}\Delta f+\frac{1}{2}G'_f(f,\mathbf{Q})\right)+\overline{\mathbf{Q}}\cdot\left((\ \mathbf{curl}\ )^2\mathbf{Q}+\frac{1}{2}G'_{\mathbf{Q}}(f,\mathbf{Q})\right)\right.\\&\left.\qquad+\int_{\partial\Sigma_\Lambda}\frac{1}{\kappa^2}\overline{f}\cdot\frac{\partial f}{\partial n}-(\ \mathbf{curl}\ \mathbf{Q}-H_0\boldsymbol{e_3})\wedge\overline{\mathbf{Q}}\cdot n\right\}.\end{aligned}$$

The last term is zero, since $(f,\mathbf{Q})$ satisfies (3.1)–(3.2). On the other hand $(f,\tilde{\mathbf{Q}})\in K_{\delta_0}$, so from the inequality of convexity we have

$$\frac{1}{2}(\overline{f},\overline{\mathbf{Q}})\cdot G''(f+t\overline{f},\mathbf{Q}+t\overline{\mathbf{Q}})\cdot(\overline{f},\overline{\mathbf{Q}})\geq c(|\overline{f}|^2+|\overline{\mathbf{Q}}|^2).$$

This implies the lemma.

4. Existence of Meissner solutions.

4.1. A priori estimates as $\boldsymbol{\kappa\to+\infty}$. In the following we consider only the functions f and $\mathbf{Q}=(Q_1,Q_2)$ defined on Ω, with curl $\mathbf{Q}=\partial_1Q_2-\partial_2Q_1$, and the associated energy defined for $(f,\mathbf{Q})\in V$, $\kappa\in(0,+\infty]$:

$$\mathcal{E}_\kappa(f,\mathbf{Q})=\int_\Omega\frac{|\nabla f|^2}{\kappa^2}+|\mathrm{curl}\ \mathbf{Q}-H_0|^2+G(f,\mathbf{Q}).$$

We recall that a critical point $(f_\kappa, \mathbf{Q}_\kappa)$ of $\mathcal{E}_\kappa$ satisfies

$$\left\{\begin{array}{l} \left.\begin{array}{l} \frac{1}{\kappa^2}\Delta f_\kappa = \frac{1}{2}G'_f(f_\kappa, \mathbf{Q}_\kappa) \\ -\,\mathbf{curl}\ (\mathrm{curl}\ \mathbf{Q}_\kappa) = \frac{1}{2}G'_{\mathbf{Q}}(f_\kappa, \mathbf{Q}_\kappa) \end{array}\right| \text{ on } \Omega, \\ \left.\begin{array}{l} \frac{\partial_n f_\kappa}{\kappa^2} = 0 \\ \mathrm{curl}\ \mathbf{Q}_\kappa = H_0 \end{array}\right| \text{ on } \partial\Omega, \end{array}\right. \tag{4.1}$$

which implies $H_\kappa = \mathrm{curl}\ \mathbf{Q}_\kappa$ satisfies

$$\left\{\begin{array}{ll} \nabla\cdot\left(\dfrac{\nabla H_\kappa}{f_\kappa^2}\right) - H_\kappa = 0 & \text{in } \Omega, \\ H_\kappa = H_0 \qquad \text{on } \partial\Omega & \end{array}\right. \tag{4.2}$$

with $\mathbf{Q}_\kappa$ given by the inverse formula

$$\mathbf{Q}_\kappa = -\frac{\mathbf{curl}\ H_\kappa}{f_\kappa^2}.$$

We are interested in the solutions of (4.1) in $\mathrm{Int}(K_{\delta_0})$. Theorem 3.1 claims the uniqueness of these solutions (when they exist). Then we have the following a priori estimates as $\kappa \to +\infty$ (which imply Proposition 2.6).

PROPOSITION 4.1. *$\forall\delta_0 > 0$, $\exists\kappa_0 = \kappa_0(\delta_0) > 0$ such that $\forall M > 1$ $\forall\theta \in (0,1)$, $\exists C_\theta = C_\theta(\delta_0, M) > 0$ such that $\forall H_0 \in [0, H^*_{\delta_0}]$ $\forall\kappa \geq \kappa_0$ for all solutions $(f_\kappa, \mathbf{Q}_\kappa) \in K^M_{\delta_0}$ of (4.1), we have*

$$\left.\begin{array}{l} |f_\kappa - f_\infty|_{C^{0,\theta}(\overline{\Omega})} \\ |\mathbf{Q}_\kappa - \mathbf{Q}_\infty|_{C^{0,\theta}(\overline{\Omega})} \\ |H_\kappa - H_\infty|_{C^{1,\theta}(\overline{\Omega})} \end{array}\right\} \leq \frac{C_\theta}{\kappa^{\frac{1}{2}-\theta}}$$

and

$$|\nabla(f_\kappa - f_\infty)|_{L^2} = O\left(\frac{1}{\kappa^{\frac{1}{2}}}\right).$$

Remark 4.2. Here C_θ depends only on θ, Ω and on $\max(M, \frac{1}{\delta_0})$.

Motivation. The essential difficulty as $\kappa \to +\infty$ is the singular perturbation of f resulting from the highest order derivatives in the equation $\frac{1}{\kappa^2}\Delta f = \frac{1}{2}G'_f$ vanishing in the limit. The limiting solution, $f = f_\infty$, does not satisfy the boundary condition ($\partial_n f_\infty \neq 0$), and there is a boundary layer of width $1/\kappa$. If (s, n) are coordinates tangential and normal to $\partial\Omega$, respectively, then rescaling $n = \bar{n}/\kappa$ in the boundary layer we find that

$$f \sim f_\infty(s,0) + \frac{1}{\kappa}\left(\partial_n f_\infty(s,0)\bar{n} + \frac{\partial_n f_\infty(s,0)}{\sqrt{2}f_\infty(s,0)}e^{-\sqrt{2}f_\infty(s,0)\bar{n}}\right)$$

there. Thus a good approximation to f both in the bulk and in the boundary layer region is

$$f_\infty(x) + \frac{1}{\kappa}\frac{\partial_n f_\infty(y)}{\sqrt{2}f_\infty(y)}e^{-\sqrt{2}f_\infty(y)\kappa d(x,\partial\Omega)}, \tag{4.3}$$

where $y \in \partial\Omega$ is such that $d(x, \partial\Omega) = |x - y|$.

Now let f_∞^κ be any approximation to f_∞ which takes into account the boundary layer, i.e., is such that

$$\partial_n f_\infty^\kappa = 0 \text{ on } \partial\Omega. \tag{4.4}$$

Write $\overline{f} = f_\kappa - f_\infty^\kappa$, $\overline{\mathbf{Q}} = \mathbf{Q}_\kappa - \mathbf{Q}_\infty$, $\overline{H} = H_\kappa - H_\infty$. Recall that $\forall H_0 \in [0, H_{\delta_0}^*]$, $(f_\infty, \mathbf{Q}_\infty) \in K_{\delta_0}$. Then under the assumption of Proposition 4.1 and (4.4) we have the following lemma.

Lemma 4.3. *For* $1 < p < +\infty$ $\forall H_0 \in [0, H_{\delta_0}^*]$, $\exists C = C(\delta_0, M, p) > 0$ *such that if* $(f_\infty^\kappa, \mathbf{Q}_\infty) \in K_{\frac{\delta_0}{2}}$ *then for every solution* $(f_\kappa, \mathbf{Q}_\kappa) \in K_{\delta_0}^M$ *of* (4.1) *we have*

$$\left.\begin{array}{l} |\overline{f}|_{L^2} \\ \frac{1}{\kappa}|\nabla\overline{f}|_{L^2} \\ |\overline{\mathbf{Q}}|_{L^2} \\ |\mathrm{curl}\overline{\mathbf{Q}}|_{\mathrm{L}^2} \end{array}\right\} \leq C\left\{\left|\frac{\Delta f_\infty^\kappa}{\kappa^2} - \frac{1}{2}[G'_f]_{f_\infty,\mathbf{Q}_\infty}^{f_\infty^\kappa,\mathbf{Q}_\infty}\right|_{L^2} + \left|\frac{1}{2}[G'_{\mathbf{Q}}]_{f_\infty,\mathbf{Q}_\infty}^{f_\infty^\kappa,\mathbf{Q}_\infty}\right|_{L^2}\right\}, \tag{4.5}$$

$$\frac{\left|\Delta\overline{f}\right|_{L^2}}{\kappa^2} \leq \left|\frac{\Delta f_\infty^\kappa}{\kappa^2} - \frac{1}{2}[G'_f]_{f_\infty,\mathbf{Q}_\infty}^{f_\infty^\kappa,\mathbf{Q}_\infty}\right|_{L^2} + \left|\frac{1}{2}[G'_f]_{f_\infty^\kappa,\mathbf{Q}_\infty}^{f_\kappa,\mathbf{Q}_\kappa}\right|_{L^2}, \tag{4.6}$$

$$\left.\begin{array}{l} \left|D^2\overline{H}\right|_{L^p} \\ \left|\nabla\overline{H}\right|_{L^p} \\ |\overline{H}|_{L^p} \end{array}\right\} \leq C(1+|\overline{H}|_{L^\infty}+|\nabla\overline{H}|_{L^\infty})\{|\overline{f}|_{L^p}+|\nabla\overline{f}|_{L^p}+|f_\infty^\kappa-f_\infty|_{L^p}+|\nabla(f_\infty^\kappa-f_\infty)|_{L^p}\}. \tag{4.7}$$

The proof of Lemma 4.3 will follow the proof of Proposition 4.1.

Remark 4.4. We note that under the assumption of Proposition 4.1 the quantities $|\frac{1}{2}[G']_{f_\infty,\mathbf{Q}_\infty}^{f_\infty^\kappa,\mathbf{Q}_\infty}|_{L^2}$ are bounded by a constant times $|f_\infty^\kappa - f_\infty|_{L^2}$. For

$$f_\infty^\kappa - f_\infty = \frac{1}{\kappa}\frac{\partial_n f_\infty(y)}{\sqrt{2}f_\infty(y)}e^{-\sqrt{2}f_\infty(y)\kappa d(x,\partial\Omega)}$$

we have $|f_\infty^\kappa - f_\infty|_{L^2} \sim C/\kappa^{\frac{3}{2}}$ which is therefore heuristically the best estimate of $|f_\kappa - f_\infty|_{L^2}$ that we can hope for.

Rather than using the exponential correction (4.3) we choose a linear function times a cut off, which makes it easier to obtain the necessary estimates. Precisely, our choice is

$$f_\infty^\kappa = f_\infty - \chi_\kappa\psi_{f_\infty},$$

where $\psi_{f_\infty}(x) = d(x,\partial\Omega)\cdot\partial_n f_\infty(y)$ where $y \in \partial\Omega$ satisfies $d(x,\partial\Omega) = |x-y|$, and where $\chi_\kappa(x) = \chi_0(\kappa d(x,\partial\Omega))$, with χ_0 a smooth nonincreasing function which satisfies $\chi_0 \geq 0$ on $[0,+\infty[$, $\chi_0 = 1$ on $[0,\frac{1}{2}]$, $\chi_0 = 0$ on $[1,+\infty[$. We assume that $\partial\Omega \in C^{2,\alpha}$; then in particular $\chi_\kappa\psi_{f_\infty} \in C^{1,\alpha}$. With this definition we have $\partial_n f_\infty^\kappa = 0$ and (compare with Remark 4.4)

$$|f_\infty^\kappa - f_\infty|_{L^2} = O\left(\frac{1}{\kappa^{\frac{3}{2}}}\right).$$

More general estimates on $f_\infty^\kappa - f_\infty$ are put together in the following lemma.

Lemma 4.5. *For $1 \leq p < +\infty$, $\theta \in (0,1)$, we have*

$$\begin{aligned}|f_\infty^\kappa - f_\infty|_{L^p} &= O\left(\frac{1}{\kappa^{1+\frac{1}{p}}}\right),\\ |\nabla(f_\infty^\kappa - f_\infty)|_{L^p} &= O\left(\frac{1}{\kappa^{\frac{1}{p}}}\right),\\ |f_\infty^\kappa - f_\infty|_{C^{0,\theta}} &= O\left(\frac{1}{\kappa^{1-\theta}}\right).\end{aligned}$$

Proof. The calculation is straightforward.

Remark 4.6. If $\partial\Omega \in C^{3,\alpha}$, then $\chi_\kappa \psi_{f_\infty} \in C^{2,\alpha}$ in which case $\Delta(f_\infty^\kappa - f_\infty)$ makes sense and we have $|\Delta(f_\infty^\kappa - f_\infty)|_{L^2} = O(\sqrt{\kappa})$.

Then we deduce the following lemma.

Lemma 4.7. *$\exists \kappa_0 = \kappa_0(\delta_0) > 0$ such that $\forall \kappa \geq \kappa_0$, for every solution $(f_\kappa, \mathbf{Q}_\kappa) \in K_{\delta_0}^M$ of* (4.1) *we have*

$$|\overline{f}|_{L^2}, |\overline{\mathbf{Q}}|_{L^2}, |curl\, \overline{\mathbf{Q}}|_{L^2} = O\left(\frac{1}{\kappa^{\frac{3}{2}}}\right), \qquad |\nabla \overline{f}|_{L^2} = O\left(\frac{1}{\kappa^{\frac{1}{2}}}\right), \tag{4.8}$$

$$|\Delta \overline{f}|_{L^2} = O(\sqrt{\kappa}), \tag{4.9}$$

and for $1 < p < +\infty$

$$|D^2 \overline{H}|_{L^p}, |\nabla \overline{H}|_{L^p}, |\overline{H}|_{L^p} = O\left(\frac{1}{\kappa^{\frac{1}{p}}} + |\overline{f}|_{L^p} + |\nabla \overline{f}|_{L^p}\right). \tag{4.10}$$

Proof of Lemma 4.7. For κ larger than some κ_0, for every $H_0 \in [0, H_{\delta_0}^*]$, we obtain $(f_\infty^\kappa, Q_\infty) \in K_{\frac{\delta_0}{2}}$, and then Lemma 4.3 can be used. The estimates (4.8) and (4.9) are directly deduced from Lemmas 4.3 and 4.5. To deduce (4.10) it is sufficient to remark that $\mathbf{Q}_\kappa = -\frac{\mathbf{curl}\ H_\kappa}{f_\kappa^2}$ and then $|\nabla H_\kappa|_{L^\infty} \leq |\mathbf{Q}_\kappa|_{L^\infty} |f_\kappa^2|_{L^\infty} \leq C$ because $(f_\kappa, \mathbf{Q}_\kappa) \in K_{\delta_0}^M$. Moreover $|\overline{H}|_{L^2} = |\text{curl}\ \overline{\mathbf{Q}}|_{L^2} = O(1/\kappa^{\frac{3}{2}})$, and therefore $|H|_{L^\infty} \leq C$, and $|\overline{H}|_{L^\infty}, |\nabla \overline{H}|_{L^\infty} \leq C$, and we deduce (4.10), which ends the proof of the lemma. □

As we have already mentioned in the introduction, we have an L^∞ constraint on the solution, while the natural spaces and estimates are L^p. Thus we need to use Gagliardo–Nirenberg inequalities (see [12]). The proof of Proposition 4.1 uses these inequalities many times, and for the convenience of the reader we recall them here for a bounded open set $\Omega \subset \mathbb{R}^n$.

Gagliardo–Nirenberg inequalities. Let $1 \leq q, r \leq +\infty$, $0 \leq j < m$, and $q' > 0$. Then

$$|D^j u|_p \leq C\{|D^m u|_r^a\, |u|_q^{1-a} + |u|_{q'}\},$$

where

$$\frac{1}{p} = \frac{j}{n} + a\left(\frac{1}{r} - \frac{m}{n}\right) + (1-a)\frac{1}{q}$$

for $\frac{j}{m} \leq a \leq 1$, and only for $\frac{j}{m} \leq a < 1$ if $m-j-\frac{n}{r} \in \mathbb{N}$ and $1 < r < +\infty$. Here $D^s u$ symbolizes the set of all partial derivatives of total order s, and for $-\infty < 1/p < +\infty$, we use the definition

$$|v|_p = \begin{cases} (\int_\Omega |v|^p)^{\frac{1}{p}} \text{ if } 0 < \frac{1}{p} < +\infty, \\ \sup_{x\in\Omega} |v(x)| \text{ if } \frac{1}{p} = 0, \\ [D^s v]_\alpha = \sup_{x,x'\in\Omega, x\neq x'} \frac{|D^s v(x') - D^s v(x)|}{|x'-x|^\alpha} \\ \qquad\qquad \text{if } -\infty < \frac{1}{p} < 0, s = [-\frac{n}{p}], \alpha = (-\frac{n}{p}) - s. \end{cases}$$

End of the proof of Proposition 4.1. We use various inequalities of Gagliardo–Nirenberg.

From the inequality

$$|\overline{f}|_{L^\infty} \leq C\{|D^2\overline{f}|_{L^2}^{\frac{1}{2}} |\overline{f}|_{L^2}^{\frac{1}{2}} + |\overline{f}|_{L^2}\};\ n = 2$$

we deduce that

$$|\overline{f}|_{L^\infty} = O\left(\kappa^{-\frac{1}{2}}\right). \tag{4.11}$$

From the inequality

$$[\overline{f}]_{2a-1} \leq C\{|D^2\overline{f}|_{L^2}^{a} |\overline{f}|_{L^2}^{1-a} + |\overline{f}|_{L^2}\};\ n = 2, \frac{1}{2} < a < 1$$

we deduce $[\overline{f}]_{2a-1} = O(\kappa^{2(a-\frac{3}{4})})$. With $\theta = 2a-1$ this implies that $\forall \theta \in (0,1), |\overline{f}|_{C^{0,\theta}} = O(1/\kappa^{\frac{1}{2}-\theta})$, and from Lemma 4.5 we deduce that

$$\forall \theta \in (0,1), |f_\kappa - f_\infty|_{C^{0,\theta}} = O\left(\frac{1}{\kappa^{\frac{1}{2}-\theta}}\right).$$

From the inequality

$$|\nabla\overline{f}|_{L^p} \leq C\{|D^2\overline{f}|_{L^2}^{a} |\overline{f}|_{L^2}^{1-a} + |\overline{f}|_{L^2}\};\ n = 2,\ \frac{1}{p} = 1-a,\ \frac{1}{2} \leq a < 1$$

we deduce $|\nabla\overline{f}|_{L^p} = O(1/\kappa^{\frac{2}{p}-\frac{1}{2}});\ 2 \leq p < +\infty$. In particular from Lemma 4.5 we deduce

$$|\nabla(f_\kappa - f_\infty)|_{L^2} = O\left(\frac{1}{\kappa^{\frac{1}{2}}}\right).$$

Moreover, from (4.11), we get $|\overline{f}|_{L^p} = O(1/\kappa^{\frac{1}{2}})$ obviously for $2 \leq p < +\infty$; and (4.10) gives $|D^2\overline{H}|_{L^p}, |\nabla\overline{H}|_{L^p}, |\overline{H}|_{L^p} = O(1/\kappa^{\frac{2}{p}-\frac{1}{2}});\ 2 \leq p < +\infty$. From the inequality

$$|\nabla\overline{H}|_{L^{\overline{p}}} \leq C\{|D^2\overline{H}|_{L^p}^{a} |\overline{H}|_{L^2}^{1-a} + |\overline{H}|_{L^2}\};\ n = 2,\ -\frac{n}{\overline{p}} = a\left(3 - \frac{2}{p}\right) - 2;$$

$$\frac{1}{2} \leq a < 1 \text{ if } p = 2;\ \frac{1}{2} \leq a \leq 1 \text{ if } 2 < p < +\infty;$$

we deduce with $\theta = 1-\frac{2}{p},\ a = 1,\ 2 < p < +\infty$, that $\forall \theta \in (0,1),\ [\nabla\overline{H}]_\theta = O(1/\kappa^{\frac{1}{2}-\theta})$. In the limit case $\theta = 0$, we have $|\nabla\overline{H}|_{L^\infty} \leq \frac{C_\eta}{\kappa^{\frac{1}{2}-\eta}}$ where η is arbitrarily small.

Therefore $\forall\theta\in(0,1), |\nabla\overline{H}|_{C^{0,\theta}}=O(1/\kappa^{\frac{1}{2}-\theta})$. In particular because $|\overline{H}|_{L^2}=O(1/\kappa^{\frac{3}{2}})$ we deduce that $|\overline{H}|_{L^\infty}=O(1/\kappa^{\frac{1}{2}-\theta})$; thus

$$\forall\theta\in(0,1),\ |\overline{H}|_{C^{1,\theta}}=O\left(\frac{1}{\kappa^{\frac{1}{2}-\theta}}\right).$$

Moreover

$$\overline{\mathbf{Q}}=\mathbf{Q}_\kappa-\mathbf{Q}_\infty=-\frac{\mathbf{curl}\ \overline{H}}{f^2}+(\ \mathbf{curl}\ H_\infty)\left(\frac{1}{f_\infty^2}-\frac{1}{f_\kappa^2}\right).$$

Thus using $|f_\infty^\kappa-f_\infty|_{C^{0,\theta}}=O(1/\kappa^{1-\theta})$ from Lemma 4.5, we deduce that

$$\forall\theta\in(0,1),\quad |\overline{\mathbf{Q}}|_{C^{0,\theta}}=O\left(\frac{1}{\kappa^{\frac{1}{2}-\theta}}\right)$$

which ends the proof of Proposition 4.1.

Proof of Lemma 4.3. From (4.1) we get

$$(4.12)\qquad \left\{\begin{array}{l} \left.\begin{array}{l}\frac{1}{\kappa^2}\Delta\overline{f}=\frac{1}{2}\{[G'_f]_{f_\infty^\kappa,\mathbf{Q}_\infty}^{f_\kappa,\mathbf{Q}_\kappa}+[G'_f]_{f_\infty,\mathbf{Q}_\infty}^{f_\infty^\kappa,\mathbf{Q}_\infty}\}-\frac{\Delta f_\infty^\kappa}{\kappa^2}\\ -\ \mathbf{curl}\ (\mathrm{curl}\ \overline{\mathbf{Q}})=\frac{1}{2}\{[G'_\mathbf{Q}]_{f_\infty^\kappa,\mathbf{Q}_\infty}^{f_\kappa,\mathbf{Q}_\kappa}+[G'_\mathbf{Q}]_{f_\infty,\mathbf{Q}_\infty}^{f_\infty^\kappa,\mathbf{Q}_\infty}\}\end{array}\right|\ \text{on }\Omega,\\ \left.\begin{array}{l}\partial_n\overline{f}=0\\ \mathrm{curl}\ \overline{\mathbf{Q}}=0\end{array}\right|\ \text{on }\partial\Omega.\end{array}\right.$$

The first equation gives the inequality (4.6). To obtain the inequality (4.5), we multiply the first equation of (4.12) (resp., the second equation of (4.12)) by $\overline{f}$ (resp., $\overline{\mathbf{Q}}$) and integrate by parts. Then we obtain

$$\int_\Omega\frac{|\nabla\overline{f}|^2}{\kappa^2}+|\mathrm{curl}\ \overline{\mathbf{Q}}|^2+\frac{1}{2}[G']_{f^\kappa,\mathbf{Q}^*}^{f,\mathbf{Q}}\cdot\left(\begin{array}{c}\overline{f}\\ \overline{\mathbf{Q}}\end{array}\right)=\int_\Omega\left(\frac{\Delta f^\kappa}{\kappa^2}-\frac{1}{2}[G'_f]_{f^*,\mathbf{Q}^*}^{f^\kappa,\mathbf{Q}^*}\right)\overline{f}-\frac{1}{2}[G'_\mathbf{Q}]_{f^*,\mathbf{Q}^*}^{f^\kappa,\mathbf{Q}^*}\cdot\overline{\mathbf{Q}}.$$

Using the inequality of convexity because $(f_\infty^\kappa,\mathbf{Q}_\infty)\in K_{\frac{\delta_0}{2}}$, we get for $c=c(\frac{\delta_0}{2})$

$$\frac{1}{2}[G']_{f_\infty^\kappa,\mathbf{Q}_\infty}^{f_\kappa,\mathbf{Q}_\kappa}\cdot\left(\begin{array}{c}\overline{f}\\ \overline{\mathbf{Q}}\end{array}\right)\geq c(\overline{f}^2+\overline{\mathbf{Q}}^2)$$

and we find the inequality (4.5).

From (4.2) we have

$$\left\{\begin{array}{l}\nabla\cdot(\frac{\nabla\overline{H}}{f_\infty^2})-\overline{H}=-g,\\ \overline{H}=0,\end{array}\right.$$

where

$$g=\nabla\cdot\left((\nabla H_\kappa)\left(\frac{1}{f_\kappa^2}-\frac{1}{(f_\infty)^2}\right)\right)=(\Delta H_\kappa)\left(\frac{1}{f_\kappa^2}-\frac{1}{(f_\infty)^2}\right)-2(\nabla H_\kappa)\left(\frac{\nabla f_\kappa}{f_\kappa^3}-\frac{\nabla f_\infty}{(f_\infty)^3}\right),$$

but from (4.2) we obtain $\Delta H_\kappa=f_\kappa^2H_\kappa+2\nabla H_\kappa\cdot\frac{\nabla f_\kappa}{f_\kappa}$ and then

$$g=\frac{H_\kappa}{(f_\infty)^2}((f_\infty)^2-f_\kappa^2)+2\frac{\nabla H_\kappa}{(f_\infty)^2}\left(\frac{\nabla f_\infty}{f_\infty}-\frac{\nabla f_\kappa}{f_\kappa}\right).$$

Then $|g|_{L^p}\leq C(|H|_{L^\infty}+|\nabla H|_{L^\infty})(|f_\kappa-f_\infty|_{L^p}+|\nabla(f_\kappa-f_\infty)|_{L^p})$. On another hand we have the standard elliptic L^p-estimate, $|\overline{H}|_{W^{2,p}}\leq C|g|_{L^p}$. Consequently we obtain the inequality (4.7), and the lemma is proved.

4.2. Local existence of the Meissner solution. We aim to find by perturbation from a known solution some solutions $(f, \mathbf{Q})$ of (4.1), that is, solutions $(f, \tilde{H})$ of the system

$$\tag{4.13} \begin{cases} \frac{1}{\kappa^2}\Delta f = \frac{1}{2}G'_f(f, -\frac{\textbf{curl } \tilde{H}}{f^2}), \\ \partial_n f = 0, \\ \nabla \cdot (\frac{\nabla \tilde{H}}{f^2}) - \tilde{H} = H_0, \\ \tilde{H} = 0, \end{cases}$$

where we have introduced the notation $\tilde{H} = H - H_0$ to obtain homogeneous boundary conditions. We will perturb using H_0. On the other hand, we are interested only in the solutions $(f, \mathbf{Q}) \in \mathrm{Int}(K_0)$, i.e., $(f, \tilde{H})$ in the open set

$$\mathcal{U} = \left\{ (f, \tilde{H}) : \text{ for } \mathbf{Q} = -\frac{\textbf{curl } \tilde{H}}{f^2},\ (f, \mathbf{Q}) \in \mathrm{Int}(K_0) \right\}.$$

Then it is natural to introduce the operator

$$\mathcal{A}(f, \tilde{H}) = \begin{cases} \frac{1}{\kappa^2}\Delta f - \frac{1}{2}G'_f(f, -\frac{\textbf{curl } \tilde{H}}{f^2}), \\ \nabla \cdot (\frac{\nabla \tilde{H}}{f^2}) - \tilde{H}. \end{cases}$$

We want to apply an inverse function theorem in Hölder spaces, which is possible because the solutions $(f, \tilde{H})$ are in fact regular (see Proposition 3.2). Define the spaces

$$X_0^{m,\alpha} = \{f, \tilde{H} \in C^{m,\alpha}(\overline{\Omega}),\ \partial_n f = 0, \tilde{H} = 0 \text{ on } \partial\Omega\},$$

$$X^{m,\alpha} = \{f, \tilde{H} \in C^{m,\alpha}(\overline{\Omega})\}.$$

We will work in the space of minimal regularity $X_0^{2,\alpha}$. In particular we assume that $\partial\Omega \in C^{2,\alpha}$. Then $\mathcal{A} : \mathcal{U}_0 = \mathcal{U} \cap X_0^{2,\alpha} \to X^{0,\alpha}$ is a regular map. We want to solve

$$\mathcal{A}(f, \tilde{H}) = \begin{pmatrix} 0 \\ H_0 \end{pmatrix}.$$

We start with the particular solution

$$\mathcal{A}(1, 0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \text{ where } (1, 0) \in \mathcal{U}_0.$$

Proof of Proposition 2.5. We have

$$D_{(f,\tilde{H})}\mathcal{A}(f, \tilde{H}) \cdot (\phi, \tilde{h}) = \begin{cases} \frac{1}{\kappa^2}\Delta\phi - 3(f^2 - \frac{|\nabla\tilde{H}|^2}{f^4} - \frac{1}{3})\phi - 2\frac{(\textbf{ curl } \tilde{H})\cdot(\textbf{ curl } \tilde{h})}{f^3}, \\ \nabla \cdot (\frac{\nabla\tilde{h}}{f^2} - 2\frac{\nabla\tilde{H}}{f^3}\phi) - \tilde{h}. \end{cases}$$

We denote $\chi = f^2 - \frac{|\nabla\tilde{H}|^2}{f^4} - \frac{1}{3}$, and we consider the equation

$$\tag{4.14} D_{(f,\tilde{H})}\mathcal{A}(f, \tilde{H}) \cdot (\phi, \tilde{h}) = \begin{cases} \phi_0 \\ \tilde{h}_0 \end{cases} \in X^{0,\alpha}.$$

We remark that if $(\phi, \tilde{h}) \in \mathrm{Ker} D\mathcal{A}(f, \tilde{H})$, then multiplying the first equation by ϕ and the second by $\tilde{h}$, and integrating by parts we obtain (with the cancellation of two terms $\mathbf{curl}\ \tilde{H} \cdot\ \mathbf{curl}\ \tilde{h} - \nabla\tilde{H} \cdot \nabla\tilde{h} = 0$)

$$\int_\Omega \frac{|\nabla\phi|^2}{\kappa^2} + 3\chi\phi^2 + \frac{|\nabla\tilde{h}|}{f^2} + \tilde{h}^2 = 0.$$

But if $(f, \tilde{H}) \in \mathcal{U}_0$ then there exists some $\eta_0 > 0$ such that $\chi > \eta_0 > 0$. Therefore $\phi = \tilde{h} = 0$ and $\mathrm{Ker} D\mathcal{A}(f, \tilde{H}) = \{0\}$.

We establish the surjectivity of $D\mathcal{A}(f, \tilde{H})$ with the help of Fredholm theory. In fact (4.14) can be written

$$\left\{ \begin{array}{l} \frac{1}{\kappa^2}\Delta\phi - 3\chi\phi - 2\frac{\nabla\tilde{H}\cdot\nabla\tilde{h}}{f^3} = \phi_0, \\ \Delta\tilde{h} - 2\frac{\nabla f\cdot\nabla\tilde{H}}{f} - 2f^2\nabla\cdot(\frac{\nabla\tilde{H}\phi}{f^3}) - f^2\tilde{h} = f^2\tilde{h}_0. \end{array} \right.$$

Here the terms of second derivatives are decoupled in ϕ and $\tilde{h}$, and we can write these equations as a compact perturbation of the identity. If we introduce (for example) the two well-known isomorphisms

$$\begin{array}{ll} A_1: & \{\phi \in C^{2,\alpha},\ \partial_n\phi = 0\} \quad \to C^{0,\alpha} \\ & \phi \longmapsto \Delta\phi - \phi \end{array},$$

$$\begin{array}{ll} A_2: & \{\tilde{h} \in C^{2,\alpha},\ \tilde{h} = 0\} \quad \to C^{0,\alpha} \\ & \tilde{h} \longmapsto \Delta\tilde{h} - \tilde{h} \end{array},$$

then (4.14) can be written

$$(\mathrm{Id} + T)(\phi, \tilde{h}) = (\phi_1, \tilde{h}_1),$$

where

$$\begin{array}{l} \phi_1 = A_1^{-1}(\kappa^2\phi_0), \\ \tilde{h}_1 = A_2^{-1}(f^2\tilde{h}_0), \end{array}$$

and

$$T(\phi, \tilde{h}) = \left\{ \begin{array}{l} A_1^{-1}((1 - 3\kappa^2\chi)\phi - 2\kappa^2\frac{\nabla\tilde{H}\cdot\nabla\tilde{h}}{f^3}), \\ A_2^{-1}((1 - f^2)\tilde{h} - 2\frac{\nabla\cdot\nabla\tilde{h}}{f} - 2f^2\nabla\cdot(\frac{\phi\nabla\tilde{H}}{f^3})). \end{array} \right.$$

Then $T: X_0^{0,\alpha} \to X_0^{0,\alpha}$ is clearly compact, and then

$$Ind(\mathrm{Id} + T) = Ind(\mathrm{Id}).$$

However, $Ind(\mathrm{Id} + T) = \mathrm{dimKer}(\mathrm{Id} + T) - codim\mathrm{Im}(\mathrm{Id} + T)$ and $Ind(\mathrm{Id}) = 0$, $\mathrm{dimKer}(\mathrm{Id}+T) = 0$, so that $\mathrm{Id}+T$ is surjective and $D\mathcal{A}(f, \tilde{H})$ is surjective. Therefore $D\mathcal{A}(f, \tilde{H}): X_0^{2,\alpha} \to X^{0,\alpha}$ is invertible and the inverse function theorem applies, which proves Proposition 2.5.

5. Global and local minimizers.

5.1. A minimizing sequence.

We start with the existence of a minimizing sequence for $\mathcal{E}_\infty$. Recall that

$$V = \{(f, \mathbf{Q}) \in L^\infty \times L^\infty,\ \nabla f \in L^2,\ \mathbf{curl}\ \mathbf{Q} \in L^2\}.$$

PROPOSITION 5.1. (i) *For any* $H_0 \geq 0$,

$$\inf_{(f,\mathbf{Q})\in V} \mathcal{E}_\infty(f, \boldsymbol{Q}) = 0.$$

The minimizing sequence can be chosen in C^∞.

(ii) *For* $H_0 > 0$ *there is no global minimizer of* $\mathcal{E}_\infty$.

Proof of Proposition 5.1. We remark that zero is the absolute minimum of $\mathcal{E}_\infty$, whose density is nonnegative. Assuming that the first part of the theorem holds, a global minimizer should satisfy almost everywhere (a.e.)

$$f^2 \left|\boldsymbol{Q}\right|^2 + \frac{(f^2-1)^2}{2} + (\text{curl } \boldsymbol{Q} - H_0)^2 = 0.$$

Then we have a.e. in Ω: $f = 1$, $\boldsymbol{Q} = 0$ and curl $\boldsymbol{Q} = H_0$. Then the distributional derivative of $\mathbf{Q}$ is zero and therefore $H_0 = 0$.

The minimizing sequence. The problem is somewhat analogous to the problem of minimizing the functional

$$\int_{-1}^{1} ((F')^2 - 1)^2 + F^2\, dx.$$

There it is well known that a minimizing sequence is furnished by the sawtooth function

$$F_n = \begin{cases} x - \frac{2j+1/2}{n} & x \in \left[\frac{2j}{n}, \frac{2j+1}{n}\right) \\ -x + \frac{2j+3/2}{n} & x \in \left(\frac{2j+1}{n}, \frac{2j+2}{n}\right] \end{cases} \quad j \in \mathbf{Z}.$$

Thus we have that $(F')^2 - 1$ is identically zero, while the derivative F_n' switches between the values -1 and $+1$ increasingly rapidly as $n \to \infty$ so that F itself stays small.

In the present case we need to satisfy curl $\boldsymbol{Q} = H_0$ while keeping $\boldsymbol{Q}$ small. If we take the second component of $\boldsymbol{Q}$ to be a sawtooth function

$$Q_2^{(n)} = H_0\left(x - \frac{j+1/2}{n}\right) \qquad x \in \left[\frac{j}{n}, \frac{j+1}{n}\right), \quad j \in \mathbf{Z},$$

then the jumps in $Q_2^{(n)}$ will create δ-function spikes in the first component at the points $x = j/n$. However, we can neutralize the effect on the energy of these spikes by simply dropping f down to zero locally.

We now construct a C^∞ minimizing sequence which is a smooth version of that described above. Let ψ be a real C^∞ function, $0 \leq \psi \leq 1$ such that

$$\psi(x) \equiv 0 \text{ for } |x| \geq 2,$$
$$\psi(x) \equiv 1 \text{ for } |x| \leq 1.$$

We introduce

$$g_n(x,y) = 1 - \sum_{k=-\infty}^{\infty} \psi\left(n^2\left(x - \frac{k}{n}\right)\right).$$

The function g_n is a sum of C^∞ functions with compact supports $\{\frac{k}{n} - \frac{2}{n^2} \leq x \leq \frac{k}{n} + \frac{2}{n^2}\}$. Now let χ_ϵ be an approximation of the identity in $\mathbb{R}$ (a nonnegative real C^∞ function with support in $(-\epsilon,\epsilon)$ and mass $\int \chi_\epsilon = 1$). Let $[x]$ denote the integral part of x (i.e., $[x] = \max\{n \in \mathbb{N}, n \leq x\}$). We define h_n as the convolution

$$h_n(x) = \left(x - \frac{[nx]}{n}\right) * \chi_{\frac{1}{2n^2}}.$$

We then define the functional

$$\boldsymbol{P_n}(x,y) = (y(h_n'(x) - 1), h_n(x)),$$

and compute curl $\boldsymbol{P_n} = h_n'(x) - (h_n'(x) - 1) = 1$. We then construct a sequence $(f_n, \boldsymbol{Q_n})$ as

$$\left\{\begin{array}{l} f_n = g_{n|\Omega}, \\ \boldsymbol{Q_n} = H_0 \boldsymbol{P}_{\boldsymbol{n}|\Omega}. \end{array}\right. \tag{5.1}$$

Let us consider the sets $B_j^n = \Omega \bigcap \bigcup_{k=-\infty}^{+\infty}\{(x,y), \{\frac{k}{n} - \frac{j}{n^2} \leq x \leq \frac{k}{n} + \frac{j}{n^2}\}$ for $j = 1, 2$. The sequence $(f_n, \boldsymbol{Q_n})$ is such that

$$\left\{\begin{array}{ll} \text{curl } \boldsymbol{Q_n} = H_0 & \text{in } \Omega, \\ 0 \leq f_n \leq 1 & \text{in } \Omega, \\ f_n = 1 & \text{in } \Omega \backslash B_2^n, \\ |\boldsymbol{Q_n}| = H_0 h_n(x) \leq \frac{H_0}{n} & \text{in } \Omega \backslash B_1^n, \\ f_n = 0 & \text{in } B_1^n. \end{array}\right.$$

We remark that $\text{meas}(B_j^n) \leq 2\frac{j}{n}(\text{diam}(\Omega))^2$. Therefore, if we estimate the energy on $\Omega \backslash B_2^n$, $B_2^n \backslash B_1^n$ and B_1^n, we obtain

$$\mathcal{E}_\infty(f_n, \boldsymbol{Q_n}) \leq \frac{C}{n}.$$

Consequently, $\inf \mathcal{E}_\infty = 0$ and $(f_n, \boldsymbol{Q_n})$ is a minimizing sequence. This completes the proof of Proposition 5.1.

From the construction above, we notice that

$$\mathcal{E}_\kappa(f_n, \boldsymbol{Q_n}) \leq \frac{C}{n} + C'\frac{n^3}{\kappa^2}.$$

Therefore, for $n = \sqrt{\kappa}$, we deduce Theorem 2.4.

5.2. Properties of local minimizers. Here we present some properties of local minimizers (see Definition 2.1).

PROPOSITION 5.2. *For $\kappa \in (0, +\infty)$, if $(f, \mathbf{Q})$ is a local minimizer in V of $\mathcal{E}_\kappa$, then $\exists \epsilon > 0$ such that either $\epsilon \leq f \leq 1 - \epsilon$ or $f \equiv 0$ or $f \equiv 1$.*

Remark 5.3. If $f \equiv 1$, then $\mathbf{Q} \equiv 0$, but curl $\mathbf{Q} = H_0$ on $\partial\Omega$, and so $H_0 = 0$.

Proof of Proposition 5.2. Let us first remark that if $(f, \mathbf{Q})$ is a global minimizer, then $\mathcal{E}_\kappa(\min(f,1),\mathbf{Q}) \leq \mathcal{E}_\kappa(f,\mathbf{Q})$ and therefore $0 \leq f \leq 1$. If $(f,\mathbf{Q})$ is a local minimizer, then let $\overline{f} = (f - (\sup_\Omega f - \epsilon))^+$. If $\sup_\Omega f > 1$, then for ϵ small enough, $\mathcal{E}_\kappa(f-\overline{f},\mathbf{Q}) < \mathcal{E}_\kappa(f,\mathbf{Q})$, which is impossible, therefore $0 \leq f \leq 1$.
Now

$$\begin{cases} \frac{1}{\kappa^2}\Delta f = f(f^2 + |\mathbf{Q}|^2 - 1), \\ f \geq 0. \end{cases}$$

Then, by the strong maximum principle, either $f \equiv 0$ or $f > 0$ on Ω. Since $\partial_n f = 0$, by the Hopf lemma we deduce that $\inf_\Omega f > 0$, and then $\inf_{\overline{\Omega}} f > 0$.

Now let $v = 1 - f$, then

$$\begin{cases} \frac{1}{\kappa^2}\Delta v - f(f+1)v = -f|\mathbf{Q}|^2 \leq 0, \\ v \geq 0. \end{cases}$$

Then, as previously, we deduce that either $v = 0$ or $\inf_{\overline{\Omega}} v > 0$, and therefore $\sup_{\overline{\Omega}} f < 1$, and the proposition is proved.

PROPOSITION 5.4. *If $(f,\mathbf{Q})$ is a critical point and there exists $\delta_0 > 0$ such that $f^2 - |\mathbf{Q}|^2 - \frac{1}{3} > \delta_0$, then $(f,\mathbf{Q})$ is a local minimizer of $\mathcal{E}_\kappa$ (for finite or infinite κ).*

Proof of Proposition 5.4. This follows immediately from Lemma 3.3.

COROLLARY 5.5. *For $\kappa = +\infty$, if $(\mathbf{Q}_\infty)^2 < \frac{1}{3} - \delta_0$ and $f_\infty > 0$, then every critical point $(f_\infty, \mathbf{Q}_\infty)$ is a local minimizer of $\mathcal{E}_\infty$.*

PROPOSITION 5.6. *For $\kappa = +\infty$, if $(f,\mathbf{Q}) \in V$ is a critical point of $\mathcal{E}_\infty$ such that*

$$\exists x_0 \in \Omega,\ |\mathbf{Q}(x_0)|^2 > \frac{1}{3} \text{ and } f(x_0) > 0$$

and $(f,\mathbf{Q})$ is continuous at $x_0 \in \Omega$, then $(f,\mathbf{Q})$ is not a local minimizer of $\mathcal{E}_\infty$.

Proof of Proposition 5.6. Let us first consider the general case $\kappa \in (0,+\infty]$. Without loss of generality we take x_0 to be the origin, and we choose the coordinates (x,y) in $\mathbb{R}^2$ such that $\mathbf{Q}(0) = |\mathbf{Q}(0)|e_x$. We consider the following function:

$$\phi(x,y) = \chi(y)\cos(\alpha x)1_{\{\alpha x \in [-\frac{\pi}{2},\frac{\pi}{2}]\}},$$

where for $\lambda, \mu > 0$,

$$\chi(y) = \begin{cases} 0 \text{ if } |y| > \frac{\lambda+\mu}{\alpha}, \\ 1 \text{ if } |y| \leq \frac{\lambda}{\alpha}, \\ \text{continuous and linear on } [-\frac{\lambda+\mu}{\alpha}, -\frac{\lambda}{\alpha}] \cup [\frac{\lambda}{\alpha}, \frac{\lambda+\mu}{\alpha}]. \end{cases}$$

Consider

$$\begin{cases} \overline{\mathbf{Q}} = \nabla\phi, \\ \overline{f} = -\Lambda\partial_x\phi. \end{cases}$$

Then

$$\mathcal{E}_\kappa(f + t\overline{f}, \mathbf{Q} + t\overline{\mathbf{Q}}) = \mathcal{E}_\kappa(f,\mathbf{Q}) + t\mathcal{E}'_\kappa(f,\mathbf{Q})\cdot(\overline{f},\overline{\mathbf{Q}}) + \frac{t^2}{2}\mathcal{E}''_\kappa(f,\mathbf{Q})\cdot(\overline{f},\overline{\mathbf{Q}})^2 + O(t^3).$$

Suppose that $(f,\mathbf{Q})$ is a local minimizer. Then $\mathcal{E}'_\kappa(f,\mathbf{Q})\cdot(\overline{f},\overline{\mathbf{Q}}) = 0$, and it remains only to examine the sign of the quantity

$$\begin{aligned} J \quad &= \tfrac{1}{2}\mathcal{E}''_\kappa(f,\mathbf{Q})\cdot(\overline{f},\overline{\mathbf{Q}})^2 \\ &= \int_\Omega \tfrac{2}{\kappa^2}|\nabla\overline{f}|^2 + 2|\ \mathbf{curl}\ \overline{\mathbf{Q}}|^2 + (f\overline{\mathbf{Q}} + 2\overline{f}\mathbf{Q})^2 + 3\overline{f}^2(f^2 - |\mathbf{Q}|^2 - \tfrac{1}{3}) \\ &= \int_\Omega \tfrac{2\Lambda^2}{\kappa^2}|\nabla\partial_x\phi|^2 - 3\Lambda^2(\partial_x\phi)^2(\tfrac{1}{3} - f^2 + |\mathbf{Q}|^2) \\ &+(\partial_x\phi)^2(f - 2\Lambda\mathbf{Q}_1)^2 + (f\partial_y\phi - 2\Lambda\mathbf{Q}_2\partial_x\phi)^2. \end{aligned}$$

Now, if we choose

$$\Lambda = \frac{f(0)}{2|\mathbf{Q}_1(0)|},$$

let $\omega = \operatorname{supp}\,\phi$, $\delta = \inf_\omega(\frac{1}{3} - f^2 + |\mathbf{Q}|^2)$, and $R = \sup_\omega(|\mathbf{Q_2}|\Lambda, |f - 2\Lambda\mathbf{Q}_1|)$, we find (using the fact that $0 \leq f \leq 1$)

$$J \leq \frac{2\Lambda^2}{\kappa^2}|\nabla\partial_x\phi|^2_{L^2} + 2|\partial_y\phi|^2_{L^2} - |\partial_x\phi|^2_{L^2}\{3\Lambda^2\delta - C'R^2\}.$$

Moreover

$$\begin{aligned} |\partial_x\phi|^2_{L^2} &= \pi\left(\lambda + \frac{\mu}{3}\right), \\ |\partial_y\phi|^2_{L^2} &= \frac{\pi}{\mu}, \\ |\nabla\partial_x\phi|^2_{L^2} &= \alpha^2(|\partial_x\phi|^2_{L^2} + |\partial_y\phi|^2_{L^2}) = \alpha^2\pi\left(\lambda + \frac{\mu}{3} + \frac{1}{\mu}\right). \end{aligned}$$

Then if we take $\lambda = 1$, $\mu = \sqrt{\alpha}$, then $|\partial_y\phi|_{L^2} \to 0$, $|\partial_x\phi|_{L^2} \to +\infty$ and $R \to 0$ as $\alpha \to +\infty$, and in particular for $\kappa = +\infty$ we find $\delta > 0$ (because $f^2 + |\mathbf{Q}|^2 = 1$). Then we obtain $J < 0$ which gives a contradiction. Consequently $(f, \mathbf{Q})$ is not a local minimizer, and Proposition 5.6 is proved.

Remark 5.7.

Large but finite $\boldsymbol{\kappa}$. For large but finite κ we can see heuristically the condition for $(f, \mathbf{Q})$ not to be a local minimizer should be

$$\left|(f^2 - |\mathbf{Q}|^2 - \frac{1}{3})^-\right|_{L^2(\omega)} \geq O\left(\frac{1}{\kappa}\right). \tag{5.2}$$

The argument goes as follows. We first remark that for finite κ if $(f, \mathbf{Q}) \in V$ is a local minimizer with $f \not\equiv 0$ and $\mathbf{Q} \not\equiv 0$, then $(f, \mathbf{Q})$ is analytic, and then if $f(x_0) = 0$ or $\mathbf{Q}(x_0) = 0$, we can chose another point x_0 arbitrarily closed to the first one such that $f(x_0) \neq 0$ and $\mathbf{Q}(x_0) \neq 0$.

Then, in the previous calculation let $L = |(f, \mathbf{Q})|_{C^1(\omega)}$ and $r^2 = (\frac{\pi}{2\alpha})^2 + (\frac{\lambda+\mu}{\alpha})^2$. Then $R^2 \leq CL^2r^2(1 + \Lambda^2)$. Then, for $\lambda = 1$ we obtain

$$\frac{J}{\pi} \leq \frac{2\Lambda^2}{\kappa^2}\alpha^2\left(1 + \frac{\mu}{3} + \frac{1}{\mu}\right) + \frac{2}{\mu} - \left(1 + \frac{\mu}{3}\right)\left\{3\Lambda^2\delta - C''\frac{L^2}{\alpha^2}(1 + \mu^2)(1 + \Lambda^2)\right\}.$$

Heuristically, if $\delta \simeq 0$, then (for κ large enough) $\Lambda \simeq \Lambda_\infty = f_\infty/(2|\mathbf{Q}_\infty|) = 1/\sqrt{2}$ (since $f^2_\infty + \mathbf{Q}^2_\infty = 1$). Then $\Lambda^2\delta \leq 1$. Let us choose $\mu = O(1/\delta)$, $\alpha = O(\kappa\sqrt{\delta})$, and $\kappa \geq L/\delta^{\frac{5}{2}}$. Then $J < 0$, so that $(f, \mathbf{Q})$ is not a local minimizer (and (5.2) is satisfied).

5.3. On the critical field at $\boldsymbol{\kappa = +\infty}$. We consider here the field H_0^* at which (for $\kappa = +\infty$) the solution $|\boldsymbol{Q}_\infty|$ first reaches $1/\sqrt{3}$. In one dimension there is an explicit solution for the infinite-κ Ginzburg–Landau equations on a halfspace $]-\infty, a]$ (see [3]), namely $H(x) = -\sqrt{2}\sinh x/\cosh^2 x$, $H(a) = H_0$, and $H' = (1-Q^2)Q$. Then $Q(a) = 1/\sqrt{3}$ when $H_0 = \sqrt{5/18} = H_0^*$, and $Q(a) = 1$ when $H_0 = 1/\sqrt{2}$. For more general domains, we have the following comparison theorem.

Proposition 5.8. *For $1/R \in \mathbb{R}$ let $B_R = \{|x| < R\}$ if $R > 0$, $B_\infty = \{(x_1, x_2), x_1 > 0\}$ if $1/R = 0$, $B_R = \{|x| > |R|\}$ if $R < 0$. Let $K_+ = \max_{x \in \partial\Omega} curv(x)$ (with $K_+ > 0$ for a disk), $K_- = \min_{x \in \partial\Omega} curv(x)$. Then*

$$H_0^*\left(B_{\frac{1}{K_-}}\right) \leq H_0^*(\Omega) \leq H_0^*\left(B_{\frac{1}{K_+}}\right).$$

In particular if Ω is convex then $H_0^*(\Omega) \geq \sqrt{5/18}$.

Proof of Proposition 5.8. This is straightforward using sub- and supersolutions.

6. Conclusion.

We have been concerned with the Meissner solution of the Ginzburg–Landau model, that is, the superconducting solution for which there are no vortices and the modulus of the order parameter is strictly positive. We have demonstrated the existence of this solution for values of the applied field H_0 less than a critical value H_0^*, and (under an additional convexity assumption) its uniqueness. We have shown that in the limit as the Ginzburg–Landau parameter $\kappa \to \infty$ the Meissner solution approaches the solution of the limiting problem formulated in [3, 1], which is a local minimizer of the limiting Ginzburg–Landau energy $\mathcal{E}_\infty$. Moreover, we have shown that the Meissner solution is only a local minimizer of the Ginzburg–Landau energy $\mathcal{E}_\kappa$, and not a global minimizer, for large enough κ.

The minimizing sequence we constructed for the energy $\mathcal{E}_\infty$ corresponds to filling the material with vortex sheets, and is in some sense comparable to the solution of the vortex density model in [5] in which the domain contains a uniform vortex density equal to the applied magnetic field. Indeed, since

$$\boldsymbol{Q} \sim \frac{1}{r}\boldsymbol{e}_\theta$$

at a vortex, a vortex may be thought of as generating a δ-function in curl $\boldsymbol{Q}$, and thus a natural way to define the Ginzburg–Landau energy for $\kappa = +\infty$ in the presence of vortices is

$$\mathcal{E}_\infty(f, \mathbf{Q}) = \int_\Omega |\text{curl } \mathbf{Q} + \omega - H_0|^2 + G(f, \mathbf{Q}),$$

where ω is the vortex density. In this case the minimizer is clearly $\boldsymbol{Q} \equiv \mathbf{0}$, $f \equiv 1$, $\omega \equiv H_0$. It is interesting that this solution is somehow found by the minimizing sequence, even when the details of the vortex cores have been ignored completely.

REFERENCES

[1] H. Berestycki, A. Bonnet, and S. J. Chapman, *A semi-elliptic system arising in the theory of type* II *superconductivity*, Comm. Appl. Nonlinear Anal., 1 (1994), pp. 1–21.

[2] H. Brezis, *Analyse fonctionnelle, theorie et applications*, Masson, Paris, 1993.

[3] S. J. Chapman, *Superheating field of type* II *superconductors*, SIAM J. Appl. Math., 55 (1995), pp. 1233–1258.

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

[5] S. J. Chapman, J. Rubinstein, and M. Schatzman, *A mean-field model of superconducting vortices*, European J. Appl. Math., 7 (1996), pp. 97–111.

[6] P. G. De Gennes, *Superconductivity of Metals and Alloys*, W. A. Benjamin, New York, 1966.

[7] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[8] R. Dautray and J.-L. Lions, *Analyse mathematique et calcul numerique pour les sciences et techniques*, Masson, Paris, 1987.

[9] V. L. Ginzburg and L. D. Landau, *On the theory of superconductivity*, Soviet Phys. JETP, 20 (1950), pp. 1064–1082.

[10] R. Monneau, *Problèmes de frontières libres, EDP elliptiques non linéaires et applications en combustion, supraconductivité et élasticité*, Doctoral Dissertation, Université Pierre et Marie Curie, Paris, 1999.

[11] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Die Grundlehrender Mathematischen Wissenschaften in Einzeldarstellungen 130, Springer-Verlag, New York, 1966.

[12] L. Nirenberg, *On elliptic partial differential equations*, Ann. Scuola Norm. Sup. Pisa (3), 13 (1959), pp. 115–162.

[13] S. Serfaty, *Stable configurations in superconductivity: uniqueness, multiplicity and vortex-nucleation*, Arch. Rational Mech. Anal. 149 (1999), pp. 329–365.

[14] S. Serfaty, *Local minimizers for the Ginzburg-Landau energy near critical magnetic field* I, Commun. Contemp. Math., 1 (1999), pp. 213–254.

[15] S. Serfaty, *Local minimizers for the Ginzburg-Landau energy near critical magnetic field* II, Commun. Contemp. Math, 1 (1999), pp. 295–333.

SIAM J. MATH. ANAL.
Vol. 31, No. 6, pp. 1396–1404

GENERIC HOPF BIFURCATION FROM LINES OF EQUILIBRIA WITHOUT PARAMETERS: II. SYSTEMS OF VISCOUS HYPERBOLIC BALANCE LAWS*

BERNOLD FIEDLER† AND STEFAN LIEBSCHER†

Abstract. We investigate viscous shock profiles of the Riemann problem for systems of hyperbolic balance laws. Even strictly hyperbolic flux terms together with a nonoscillating kinetic part can lead to oscillating viscous shock profiles. They appear near a Hopf-like bifurcation point of the traveling wave equation.

Key words. viscous profiles, oscillating viscous shocks, Riemann problem, hyperbolic balance laws

AMS subject classifications. 35L67, 34C23, 34C37

PII. S0036141098341721

1. Introduction.

Searching for viscous shock profiles of the Riemann problem, we consider systems of hyperbolic balance laws of the form

$$u_t + f(u)_x = \epsilon^{-1} g(u) + \epsilon\delta u_{xx}, \tag{1.1}$$

with $u = (u_0, u_1, \ldots, u_N) \in \mathbb{R}^{N+1}, f \in C^3, g \in C^2, \delta > 0$, and with real time t and space x. We assume strict hyperbolicity, that is, the Jacobian $A(u) = f'(u)$ possesses simple real distinct eigenvalues

$$\sigma_0, \sigma_1, \ldots, \sigma_N \in \text{spec } A(u) \tag{1.2}$$

The case of conservation laws, $g \equiv 0$, has been studied extensively. See, for example, [8] for a background. Viscous profiles are traveling wave solutions of the form

$$u = u\left(\frac{x - st}{\epsilon}\right) \tag{1.3}$$

with wave speed s. Here (1.3) provides a solution of (1.1) if

$$-s\dot{u} + A(u)\dot{u} = g(u) + \delta\ddot{u}. \tag{1.4}$$

Here $A(u) = f'(u)$ denotes the Jacobian and $\dot{} = \frac{d}{d\tau}$ with $\tau = (x - st)/\epsilon$. Note that (1.4) is independent of $\epsilon > 0$. Any solution of system (1.4) for which

$$\lim_{\tau \to \pm\infty} u(\tau) = u^{\pm} \tag{1.5}$$

exists gives rise, for $\epsilon \searrow 0$, to a solution of the Riemann problem of (1.1) with values $u = u^{\pm}$ connected by a shock traveling with shock speed s. We call solutions $u(\cdot)$ of (1.4), (1.5) *viscous profiles.*

*Received by the editors July 7, 1998; accepted for publication (in revised form) November 17, 1999; published electronically June 15, 2000. This work was supported by the DFG-Schwerpunkt Analysis und Numerik von Erhaltungsgleichungen and with funds provided by the National Science Foundation at IMA, University of Minnesota, Minneapolis.

http://www.siam.org/journals/sima/31-6/34172.html

†Institut für Mathematik I, Freie Universität Berlin, Arnimallee 2-6, 14195 Berlin, Germany (fiedler@math.fu-berlin.de, liebsch@math.fu-berlin.de).

We rewrite the viscous profile equation (1.4) as a second order system

$$\begin{aligned} \dot{u} &= v, \\ \delta\dot{v} &= -g(u) + (A(u) - s)v. \end{aligned} \tag{1.6}$$

Note that any viscous profile must satisfy

$$g(u^{\pm}) \;=\; 0. \tag{1.7}$$

In other words, the asymptotic states $u^{\pm}$ must be equilibria of the reaction term $g(u)$. In the conservation law case, $g \equiv 0$, this condition does not impose any constraint on the Riemann values $u^{\pm}$. In the other extreme of a reaction term g with unique equilibrium, we obtain $u^{+} = u^{-}$ and traveling shock profiles of Riemann type do not exist. In the case of one conservation law mixed with N balance laws, one expects curves of equilibria $g(u^{\pm}) = 0$. For example, reaction terms $g(u)$ typically depend on concentrations or temperature, but not on velocity.

Addressing a simple case, for demonstration purposes, we therefore assume that the u_0-component does not contribute to the reaction terms and still all that reaction components vanish, say at $u = 0$. Specifically, we assume throughout this paper that

$$g = g(u_1, \ldots, u_N) = \begin{pmatrix} g_0 \\ g_1 \\ \vdots \\ g_N \end{pmatrix} \tag{1.8}$$

is independent of u_0 and satisfies

$$g(0) = 0. \tag{1.9}$$

This gives rise to a line of equilibria

$$u_0 \in \mathbb{R}, \quad u_1 = \cdots = u_N = 0, \quad v = 0 \tag{1.10}$$

of our viscous profile system (1.6).

The asymptotic behavior of viscous profiles $u(\tau)$ for $\tau \to \pm\infty$ depends on the linearization L of (1.6) at $u = u_{\pm}$, $v = 0$. In block matrix notation corresponding to coordinates (u, v) we have

$$L = \begin{pmatrix} 0 & \mathrm{id} \\ -\delta^{-1} g' & \delta^{-1}(A - s). \end{pmatrix} \tag{1.11}$$

Here $A = A(u)$, and $g' = g'(u)$ describes the Jacobi matrix of the reaction term g at $u = u^{\pm}$. In (1.11) we write s rather than $s \cdot \mathrm{id}$ for brevity. In the case $g \equiv 0$ of pure conservation laws, the linearization L possesses an $(N+1)$-dimensional kernel corresponding to the then arbitrary choice of the equilibrium $u \in \mathbb{R}^{N+1}$, $v = 0$. Normal hyperbolicity of this family of equilibria, in the sense of dynamical systems [5], [2], [9], is ensured for wave speeds s not in the spectrum of the strictly hyperbolic Jacobian $A(u)$:

$$s \notin \mathrm{spec}\; A(u) = \{\sigma_0, \ldots, \sigma_N\}. \tag{1.12}$$

Indeed, (1.12) ensures that additional zeros do not arise in the real spectrum

$$\mathrm{spec}\; L = \{0\} \cup \delta^{-1}\,\mathrm{spec}\,(A(u) - s). \tag{1.13}$$

In the present paper we investigate the failure of normal hyperbolicity of L along the line of equilibria $u = (u_0, 0, \ldots, 0), v = 0$, given by (1.10). Although our method applies in complete generality, we present just a simple specific example for which purely imaginary eigenvalues of L arise when $\delta > 0$ is fixed small enough. Specifically, we consider three-dimensional systems, $N = 2$, satisfying

$$\begin{aligned} A(u_0,0,0) &= A_0 + u_0 \cdot A_1, \\ A_0 &= \begin{pmatrix} \alpha & & \\ & 1 & \\ & & -1 \end{pmatrix}, \quad \alpha \neq 0, \quad A_1 \text{ symmetric}, \\ g'(0) &= \begin{pmatrix} 0 & & \\ & \gamma & 1 \\ & 1 & \gamma \end{pmatrix}, \quad |\gamma| < 1, \end{aligned} \tag{1.14}$$

with omitted entries being zero. Note that these data can arise from flux functions f which are gradient vector fields, still giving rise to purely imaginary eigenvalues. At the end of this paper we present a specific example where the reaction terms $\dot{u} = g(u)$ *alone*, likewise, do not support even transient oscillatory behavior; see (3.12). The interaction of flux and reaction, in contrast, is able to produce purely imaginary eigenvalues of the linearization L, as follows.

PROPOSITION 1.1. *Consider the linearization $L = L(u_0)$ along the line $u_0 \in \mathbb{R}, u_1 = u_2 = 0$ of equilibria of the viscous profile system* (1.6) *in $\mathbb{R}^3$; see* (1.11). *Assume* (1.14) *holds.*

For $\delta \searrow 0$ and small $|s|, |u_0|$, the spectrum of L then decouples into two parts:

(i) *an unboundedly growing part* $\operatorname{spec}_\infty(L) = \delta^{-1}\operatorname{spec}(A - s) + O(1)$,

(ii) *a bounded part* $\operatorname{spec}_{bd}(L) = \operatorname{spec}((A-s)^{-1}g') + O(\delta)$.

Here A, g are evaluated at $u = (u_0, 0, 0)$.

For $\delta = 0, s = 0, |\gamma| < 1$, the bounded part $\operatorname{spec}_{bd}(L)$ *at $u_0 = 0$ limits onto simple eigenvalues $\mu_0 \in \{0, \pm i\omega_0\}$, $\omega_0 = \sqrt{1-\gamma^2}$ with eigenvectors $\left(\begin{smallmatrix} \tilde{u} \\ \tilde{v} \end{smallmatrix}\right)$ given by $\tilde{v} = \mu_0\tilde{u}$ and*

$$\begin{aligned} \tilde{u} &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad \textit{for } \mu_0 = 0, \\ \tilde{u} &= \begin{pmatrix} 0 \\ -\gamma - i\omega_0 \\ 1 \end{pmatrix} \qquad \textit{for } \mu_0 = +i\omega_0. \end{aligned} \tag{1.15}$$

Proof. Regular perturbation theory applies to the scaled block matrix

$$\delta L = \begin{pmatrix} 0 & 0 \\ -g' & A - s \end{pmatrix} + \delta \begin{pmatrix} 0 & \mathrm{id} \\ 0 & 0 \end{pmatrix}, \tag{1.16}$$

which becomes lower triangular for $\delta = 0$. This provides us with the unbounded part $\operatorname{spec}_\infty(L)$ of the spectrum, generated in v-space alone with $u = 0$.

Moreover, δL possesses three-dimensional kernel, at $\delta = 0$, given by

$$g'u = (A - s)v. \tag{1.17}$$

On this kernel, the eigenvalue problem for L reduces to

$$\mu_0 u \;=\; v \;=\; (A-s)^{-1} g' u. \tag{1.18}$$

The characteristic polynomial of (1.18) at $u_0 = 0$ is given by

$$p_0(\mu) = \left(\mu^2 - \frac{2\gamma s}{1-s^2}\mu + \frac{1-\gamma^2}{1-s^2}\right)\mu. \tag{1.19}$$

Direct calculation completes the proof. □

Note that our choices of A_0 and $g'(0)$ are normalized, such that the bifurcation occurs at a shock speed $s = 0$. For more general systems the Hopf point may occur at nonzero values of the shock speed parameter.

For explicit calculations here and below, we have used and recommend assistance by symbolic packages like Mathematica, Maple, etc.

Bifurcations from lines of equilibria in absence of parameters have been investigated in [6], [3] from a theoretical view point. We briefly recall that result for convenience. Consider C^5 vector fields

$$\dot{\mathbf{u}} = F(\mathbf{u}) \tag{1.20}$$

with $\mathbf{u} = (u_0, u_1, \ldots, u_n) \in \mathbb{R}^{n+1}$. We assume a line of equilibria

$$0 = F(u_0, 0, \ldots, 0) \tag{1.21}$$

along the u_0-axis. At $u_0 = 0$, we assume the Jacobi matrix $F'(u_0, 0, \ldots, 0)$ to be hyperbolic, except for a trivial kernel vector along the u_0-axis and a complex conjugate pair of simple, purely imaginary, nonzero eigenvalues $\mu(u_0), \overline{\mu(u_0)}$ crossing the imaginary axis transversely as u_0 increases through $u_0 = 0$:

$$\begin{aligned} \mu(0) &= i\omega(0), \quad \omega(0) > 0, \\ \operatorname{Re}\mu'(0) &\neq 0. \end{aligned} \tag{1.22}$$

Let Z be the two-dimensional real eigenspace of $F'(0)$ associated to $\pm i\omega(0)$. By Δ_Z we denote the Laplacian with respect to variations of u in the eigenspace Z. Coordinates in Z are chosen as coefficients of the real and imaginary parts of the complex eigenvector associated to $i\omega(0)$. Note that the linearization acts as a rotation with respect to these not necessarily orthogonal coordinates. Let P_0 be the one-dimensional eigenprojection onto the trivial kernel along the u_0-axis. Our final nondegeneracy assumption then reads

$$\Delta_Z P_0 F(0) \neq 0. \tag{1.23}$$

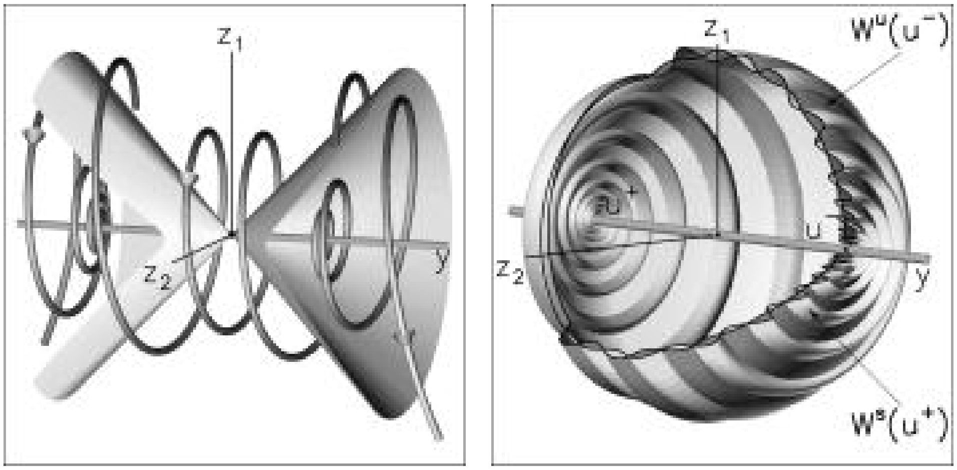
Fixing orientation along the positive u_0-axis, we can consider $\Delta_Z P_0 F(0)$ as a real number. Depending on the sign

$$\eta := \operatorname{sign}\,(\operatorname{Re}\mu'(0)) \;\cdot\; \operatorname{sign}\,(\Delta_Z P_0 F(0)), \tag{1.24}$$

we call the "bifurcation" point $u_0 = 0$ *elliptic* if $\eta = -1$ and *hyperbolic* for $\eta = +1$.

The following result from [3] investigates the qualitative behavior of solutions in a normally hyperbolic three-dimensional center manifold to $\mathbf{u} = 0$.

The results for the hyperbolic case $\eta = +1$ are based on normal form theory and a spherical blow-up construction inside the center manifold. The elliptic case $\eta = -1$



Case (a): hyperbolic, $\eta = +1$. Case (b): elliptic, $\eta = -1$.

FIG. 1.1. *Dynamics near Hopf bifurcation from lines of equilibria.*

is based on Neishtadt's theorem on exponential elimination of rapidly rotating phases [7]. For a related application to binary oscillators in discretized systems of hyperbolic balance laws see [4]. For an application to square rings of additively coupled oscillators see [1].

THEOREM 1.2. *Let assumptions* (1.21)–(1.23) *hold for the* C^5 *vector field* $\dot{\mathbf{u}} = F(\mathbf{u})$. *Then the following holds true in a neighborhood* U *of* $\mathbf{u} = 0$ *within a three-dimensional center manifold to* $\mathbf{u} = 0$.

In the hyperbolic case, $\eta = +1$, *all nonequilibrium trajectories leave the neighborhood* U *in positive or negative time direction (possibly both). The stable and unstable sets of* $\mathbf{u} = 0$, *respectively, form cones around the positive/negative* u_0-*axis, with asymptotically elliptic cross section near their tips at* $\mathbf{u} = 0$. *These cones separate regions with different convergence behavior. See Figure* 1.1 (a).

In the elliptic case all nonequilibrium trajectories starting in U *are heteroclinic between equilibria* $\mathbf{u}^{\pm} = (u_0^{\pm}, 0, \ldots, 0)$ *on opposite sides of* $u_0 = 0$. *If* $F(\mathbf{u})$ *is real analytic near* $\mathbf{u} = 0$, *then the two-dimensional strong stable and strong unstable manifolds of* $\mathbf{u}^{\pm}$ *within the center manifold intersect at an angle which possesses an exponentially small upper bound in terms of* $|\mathbf{u}^{\pm}|$. *See Figure* 1.1 (b).

In the present paper, we apply Theorem 1.2 to the problem of zero speed viscous profiles of systems of hyperbolic balance laws near Hopf points as in Proposition 1.1. Nonzero shock speeds can be treated completely analogously, absorbing them into the flux term.

THEOREM 1.3. *Consider the problem* (1.5)–(1.7) *of finding viscous profiles with shock speed* $s = 0$ *to hyperbolic balance laws* (1.1). *Let assumptions* (1.8)–(1.10), (1.14) *hold, so that a pair of purely imaginary simple eigenvalues occurs for the linearization* L, *in the limit* $\delta \to 0$.

Then there exist nonlinearities $A(u) = f'(u)$ *and* $g(u)$, *compatible with the above assumptions, such that the assumptions and conclusions of Theorem* 1.2 *are valid for the viscous profile system* (1.6). *Both the elliptic and the hyperbolic cases occur; see Figure* 1.1.

Since both conditions are open, the results persist, in particular, for small nonzero shock speeds s, *even when* f, g *remain fixed.*

Specific choices of flux $f(u)$ and reaction terms $g(u)$ are presented in Corollary 3.3; see (3.12). In the elliptic case $\eta = -1$, we nevertheless observe (at least) pairs of weak shocks with oscillatory tails, connecting u_- and u_+. In the hyperbolic case $\eta = +1$, viscous profiles leave the neighborhood U and thus represent large shocks. At $u_0 = 0$, their profiles change discontinuously and the role of the u_0-axis switches from providing the left to providing the right asymptotic state with oscillatory tail.

This paper is organized as follows. In section 2 we check transversality condition (1.22) for the purely imaginary eigenvalues. We also compute an expansion in terms of δ for the eigenprojection P_0 onto the trivial kernel along the u_0-axis. In section 3, we check nondegeneracy condition (1.23) for $\Delta_Z P_0 F(0)$, in the limit $\delta \searrow 0$, completing the proof of Theorem 1.3 by reduction to Theorem 1.2.

2. Linearization and transverse eigenvalue crossing. In this section we continue our analysis of the linearization

$$L^\delta(u_0) = \begin{pmatrix} 0 & \mathrm{id} \\ -\delta^{-1} g' & \delta^{-1} A \end{pmatrix}, \tag{2.1}$$

with $A = A(u)$, $g' = g'(u)$ evaluated along the line of equilibria $u = (u_0, 0, 0)$. See (1.11) with $s = 0$ and Proposition 1.1. In the limit $\delta \searrow 0$, we address the issue of transverse crossing of purely imaginary eigenvalues in Lemma 2.1. In Lemma 2.2, we explicitly compute the one-dimensional eigenprojection P_0^δ onto the trivial kernel of $L^\delta(0)$.

Throughout this section we fix the notation

$$A(u_0, 0, 0) = A_0 + u_0 A_1 = \begin{pmatrix} \alpha & & \\ & 1 & \\ & & -1 \end{pmatrix} + u_0 \cdot (a^1_{jk})_{0 \le j,k \le 2}, \tag{2.2}$$

with $a^1_{jk} = a^1_{kj}$ symmetric; see (1.14). We also assume that the linearized reaction term

$$g'(u_0, 0, 0) = g'(0) = \begin{pmatrix} 0 & & \\ & \gamma & 1 \\ & 1 & \gamma \end{pmatrix}, \qquad |\gamma| < 1, \tag{2.3}$$

which is independent of u_0 by assumption (1.8), possesses a vanishing g_0-component.

By Proposition 1.1, purely imaginary eigenvalues of $L^\delta(u_0)$ arise from an $O(\delta)$ perturbation of the matrix

$$A^{-1} g' = (A_0 + u_0 A_1)^{-1} g'(0) \tag{2.4}$$

with spectrum $\mathrm{spec}_{bd}(L)$. Let

$$\mu(u_0), \quad \bar\mu(u_0) \tag{2.5}$$

denote the continuation of the simple, purely imaginary eigenvalues

$$\mu(0) = i\omega_0, \quad \bar\mu(0) = -i\omega_0$$

with u_0-derivatives $\mu'(u_0), \bar\mu'(u_0)$.

Lemma 2.1. *In the above setting and notation,*

$$\mathrm{Re}\, \mu'(0) = -\frac{\gamma}{2}(a^1_{11} + a^1_{22}) + a^1_{12}. \tag{2.6}$$

Proof. Since the unit vector e_0 in u_0-direction is a trivial kernel vector of $g'(0)$ and since the remaining eigenvalues of $A^{-1}g'$ remain conjugate complex for small $|u_0|$, we have

$$\text{Re}\ \mu(u_0) = \frac{1}{2}\ \text{trace}\ (A^{-1}g'). \tag{2.7}$$

In particular, trace $A_0^{-1}g' = 0$. With the u_0-expansion

$$A^{-1} = (A_0 + u_0A_1)^{-1} = A_0^{-1} - u_0A_0^{-1}A_1A_0^{-1} + \cdots \tag{2.8}$$

we immediately obtain

$$\text{Re}\ \mu'(u_0) = -\frac{1}{2}\ \text{trace}\,(A_0^{-1}A_1A_0^{-1}g'). \tag{2.9}$$

Inserting A_0, A_1, g' proves the lemma. □

By regular perturbation of $\text{spec}_{bd}\,(L)$, the result $\text{Re}\ \mu'(0) \neq 0$ of Lemma 2.1 extends to small positive δ.

We now turn to an expansion for the eigenprojection P_0^δ onto the one-dimensional kernel of the 6×6-matrix $L^\delta(u_0)$ at $u_0 = 0$; see (2.1)–(2.3). Aligning the notations of Proposition 1.1 and of Theorem 1.2, we decompose

$$\mathbf{u} = (u, v) \in \mathbb{R}^6 = \mathbb{R}^3 \times \mathbb{R}^3.$$

Again, $e_0^T = (1, 0, 0)$ denotes the first unit vector in $\mathbb{R}^3$ and $\mathbf{e}_0^T = (e_0^T, 0)$ the first unit vector in $\mathbb{R}^6$.

Lemma 2.2. *In the above setting and notation*

$$\begin{aligned} P_0^\delta &= \mathbf{e}_0 \cdot \mathbf{e}_\delta^T, \qquad \textit{with} \\ \mathbf{e}_\delta^T &= (1 + (\tfrac{\delta}{\alpha})^2)^{-1/2}(e_0^T, -\tfrac{\delta}{\alpha}e_0^T). \end{aligned} \tag{2.10}$$

Proof. Kernel and cokernel of $L^\delta(u_0)$ are one-dimensional, corresponding to the simple zero eigenvalue of $L^\delta(u_0)$. Obviously

$$\ker L^\delta(u_0) = \mathbf{e}_0, \tag{2.11}$$

because $g'(u_0, 0, 0)e_0 = 0$. At $u_0 = 0$, the cokernel of $L^\delta(u_0)$ is given by

$$0 = \mathbf{e}_\delta^T \cdot \begin{pmatrix} 0 & \text{id} \\ -\delta^{-1}g' & \delta^{-1}A_0 \end{pmatrix}. \tag{2.12}$$

Inserting A_0 from (2.2) and g' from (2.3) proves the lemma. □

3. Higher order nondegeneracy. In this section we complete the proof of Theorem 1.3. In view of Theorem 1.2, we have already checked transverse crossing of purely imaginary eigenvalues, assumption (1.22), in Lemma 2.1. Letting

$$F(\mathbf{u}) = F(u, v) = \begin{pmatrix} v \\ -\delta^{-1}g(u) + \delta^{-1}A(u)v \end{pmatrix} \tag{3.1}$$

it remains to check the nondegeneracy assumption $\Delta_Z P_0 F \neq 0$; see (1.23). In Lemma 3.1, we check this assumption in the limit $\delta \searrow 0$. In Corollary 3.2, we provide explicit expressions for the type determining sign $\eta = \pm 1$ defined in (1.24). In particular, we show in Corollary 3.3 that both the hyperbolic case $\eta = +1$ and the elliptic case $\eta = -1$ can be realized by our nonlinear hyperbolic balance laws, even with gradient flux terms. This then completes the proof of Theorem 1.3.

To check nondegeneracy condition (1.23) on $\Delta_Z P_0 F$ in the limit $\delta \searrow 0$, we use the following notation. By transverse eigenvalue crossing at $\delta = 0$, Lemma 2.1, we also obtain purely imaginary eigenvalues $\pm i\omega^\delta$ at equilibria $\mathbf{u}^\delta = (u_0^\delta, 0, \ldots, 0) = (u^\delta, 0)$ on the u_0-axis, for small $\delta > 0$. Let Z^δ denote the corresponding eigenspace. We recall our expression for the eigenprojection P_0^δ onto the trivial kernel,

$$P_0^\delta = (1 + (\tfrac{\delta}{\alpha})^2)^{-1/2}\, \mathbf{e}_0 \cdot (e_0^T, -\tfrac{\delta}{\alpha} e_0^T) \tag{3.2}$$

with $\mathbf{e}_0 = (e_0^T, 0)^T$, see Lemma 2.2. Note that $\mathbf{u}^\delta, \omega^\delta, P_0^\delta$, and Z^δ vary differentiably with δ.

Lemma 3.1. *In the above setting and notation we have*

$$\Delta_{Z^\delta} P_0^\delta F(\mathbf{u}^\delta) = (1 + (\tfrac{\delta}{\alpha})^2)^{-1/2}\, \frac{1}{\alpha} g_0''(u^\delta)[\tilde{u}^\delta, \bar{\tilde{u}}^\delta]\, \mathbf{e}_0 \tag{3.3}$$

at the Hopf point $\mathbf{u}^\delta = (u^\delta, 0)$ with complex eigenvector $(\tilde{u}^\delta, \tilde{v}^\delta)$ of $i\omega^\delta$.

Consider, in particular, quadratic forms $g_0''(0)$, which are strictly positive/negative definite on (u_1, u_2)-space, with $\Gamma = \pm 1$ indicating the sign of definiteness. Then

$$\operatorname{sign} \Delta_{Z^\delta} P_0^\delta F(\mathbf{u}^\delta) = \Gamma \cdot \operatorname{sign} \alpha \tag{3.4}$$

for all small $\delta > 0$.

Proof. By Lemma 2.2 we have

$$(1 + (\tfrac{\delta}{\alpha})^2)^{1/2} P_0^\delta = \mathbf{e}_0 \cdot \mathbf{e}_0^T - \tfrac{\delta}{\alpha} \mathbf{e}_0 \cdot (0, e_0^T). \tag{3.5}$$

The explicit form (3.1) of the nonlinearity F implies

$$\Delta_{Z^\delta} P_0^0 F(\mathbf{u}^\delta) = \mathbf{e}_0 \Delta_{Z^\delta} v_0 = 0 \tag{3.6}$$

on any subspace Z^δ and for any $\mathbf{u}^\delta$, simply because the u-component of F is linear. With P_0^δ instead of P_0^0 we obtain more generally

$$\begin{aligned} (1 + (\tfrac{\delta}{\alpha})^2)^{1/2} \mathbf{e}_0^T \Delta_{Z^\delta} P_0^\delta F(\mathbf{u}) &= -\Delta_{Z^\delta} \left(0, -\tfrac{\delta}{\alpha} e_0^T \delta^{-1}(-g(u) + A(u)v)\right) \\ &= -\tfrac{1}{\alpha} \Delta_{Z^\delta} \left(-g_0(u) + (A(u)v)_0\right). \end{aligned} \tag{3.7}$$

Here $(A(u)v)_0$ denotes the zero-component of $A(u)v$. We treat this term first, using the notation

$$\tilde{\mathbf{u}}^\delta = \begin{pmatrix} \tilde{u}^\delta \\ \tilde{v}^\delta \end{pmatrix} \tag{3.8}$$

for the complex eigenvector of the purely imaginary Hopf eigenvalue $\mu^\delta = i\omega^\delta$ at $u = u^\delta, v = 0$. Then $Z^\delta = \operatorname{span}\{\mathrm{Re}\, \tilde{\mathbf{u}}^\delta, \mathrm{Im}\, \tilde{\mathbf{u}}^\delta\}$. Denoting by $\Delta_\beta = \partial_{\beta_1}^2 + \partial_{\beta_2}^2$ the standard Laplacian, evaluated at $\beta = 0$, and inserting $\tilde{v}^\delta = \mu^\delta \tilde{u}^\delta$ yields

$$\begin{aligned} \Delta_{Z^\delta}(A(u)v)_0 &= \Delta_\beta\big(A(u^\delta + \beta_1 \mathrm{Re}\, \tilde{u}^\delta + \beta_2 \mathrm{Im}\, \tilde{u}^\delta)\ \ (\beta_1\, \mathrm{Re}\, \tilde{v}^\delta + \beta_2 \mathrm{Im}\, \tilde{v}^\delta)\big)_0 \\ &= 2\left((A'(u^\delta)\mathrm{Re}\, \tilde{u}^\delta)\mathrm{Re}\, \tilde{v}^\delta + (A'(u^\delta)\mathrm{Im}\, \tilde{u}^\delta)\mathrm{Im}\, \tilde{v}^\delta\right)_0 \\ &= 2\, \mathrm{Re}\, \left((A'(u^\delta)\tilde{u}^\delta)\bar{\tilde{v}}^\delta\right)_0 \\ &= 2\, \mathrm{Re}\, (\bar{\mu}^\delta)\, \left(f''(u^\delta)[\tilde{u}^\delta, \bar{\tilde{u}}^\delta]\right)_0 \\ &= 2\, \mathrm{Re}\, (\mu^\delta)\, f_0''(u^\delta)[\tilde{u}^\delta, \bar{\tilde{u}}^\delta] = 0 \end{aligned} \tag{3.9}$$

all along the Hopf curve $u = u^\delta, v = 0$. Here we have used $A(u) = f'(u)$ for the flux function and the fact that the Hessian matrix $f_0^{''}(0)$ is symmetric.

Therefore, we can conclude from (3.7), (3.9) that

$$\left(1 + \left(\tfrac{\delta}{\alpha}\right)^2\right)^{1/2} \mathbf{e}_0^T \Delta_{Z^\delta} P_0^\delta F(\mathbf{u}^\delta) = \frac{1}{\alpha}\Delta_{Z^\delta} g_0(u) = \frac{1}{\alpha} g_0''(u^\delta)[\tilde{u}^\delta, \bar{\tilde{u}}^\delta]. \tag{3.10}$$

This proves (3.3) and the lemma. □

Corollary 3.2. *Combining Lemmas* 2.1 *and* 3.1*, the* sign $\eta = \pm 1$ *distinguishing elliptic from hyperbolic Hopf bifurcation along our line of equilibria is given explicitly by*

$$\begin{aligned} \eta = &\quad \text{sign Re } \mu'(0) \cdot \text{sign}\, \Delta_Z P_0 F(0) \\ = &\quad \text{sign}(a_{12}^1 - \tfrac{\gamma}{2}(a_{11}^1 + a_{22}^1)) \cdot \text{sign}\, \alpha \cdot \Gamma, \end{aligned} \tag{3.11}$$

for $\delta > 0$ small enough. Here derivatives are evaluated at $u = 0$ and are assumed to be chosen such that $\eta \neq 0$. The sign $\Gamma = \pm 1$ *indicates positive/negative definiteness of $g_0''(0)$ on (u_1, u_2)-space. Obviously, both signs of η can be realized.*

Corollary 3.3. *Theorems* 1.2, 1.3 *hold true for $\eta = \pm 1$ with the following specific choices of a gradient flux term $f(u) = \nabla\Phi(u)$ and a reaction term $g(u)$:*

$$\begin{aligned} g(u) &= \begin{pmatrix} u_1^2 & + & u_2^2 \\ -\frac{1}{2}u_1 & + & u_2 \\ u_1 & - & \frac{1}{2}u_2 \end{pmatrix}, \\ \Phi(u) &= u_0^2 + \tfrac{1}{2}(u_1^2 - u_2^2) + \eta u_0 u_1^2. \end{aligned} \tag{3.12}$$

These choices correspond to $\alpha = 2, \gamma = -\frac{1}{2}, \Gamma = +1$.

Acknowledgments. Helpful discussions with Bob Pego and Heinrich Freistühler concerning the set-up of viscous profiles for hyperbolic balance laws are gratefully acknowledged. This work was supported by the Deutsche Forschungsgemeinschaft, Schwerpunkt "Analysis und Numerik von Erhaltungsgleichungen," and was completed during a fruitful and enjoyable research visit of one of the authors at the Institute of Mathematics and Its Applications, University of Minnesota, Minneapolis, supported with funds provided by the National Science Foundation.

REFERENCES

[1] J. Alexander and B. Fiedler, *Stable and unstable decoupling in squares of additively coupled oscillators*, in preparation.

[2] N. Fenichel, *Asymptotic stability with rate conditions,* II, Indiana Univ. Math. J., 26 (1977), pp. 81–93.

[3] B. Fiedler, S. Liebscher, and J. Alexander, *Generic Hopf bifurcation from lines of equilibria without parameters:* I. *Theory*, J. Differential Equations, to appear.

[4] B. Fiedler, S. Liebscher, and J. Alexander, *Generic Hopf bifurcation from lines of equilibria without parameters:* III. *Binary oscillations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., to appear.

[5] M. Hirsch, C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.

[6] S. Liebscher, *Stabilität von Entkopplungsphänomenen in Systemen gekoppelter symmetrischer Oszillatoren*, Diploma Thesis, Free University Berlin, Berlin, Germany, 1997.

[7] A. Neishtadt, *On the separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech., 48 (1984), pp. 134–139.

[8] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[9] S. Wiggins, *Normally Hyperbolic Invariant Manifolds in Dynamical Systems*, Springer-Verlag, New York, 1994.